NetGAN:通过随机游走生成图阅读报告

葛超

文章提出现实世界的网络存在的特征，质疑已建立的模型的有效性。引导我们进入一个问题：“我们如何定义一个模型来捕捉真实图的所有本质的（潜在的未知）性质？”

而目前为止，没有任何有关真实世界网络的GAN结构的作品被提出。作者提出了两种相关的工作进行比较：Liu的工作和Tavakoli的工作。

1，Liu的工作：提出了一种用来学习子图的拓扑性质的GAN结构

2，Tavakoli的工作：应用GAN的训练方式直接生成表示图的邻接矩阵来生成图

这两种都是不可行的：由于这两种模型都是使用整个邻接矩阵的数据，包括了0元素。这些0元素在作者看来是不需要放在计算过程的。将所有的邻接矩阵进行考虑是低效率的。作者借此提出了NetGAN:首先NetGAN是在随机漫步上进行的—仅仅考虑了邻接矩阵中的非零的部分，这样也可以很方便的处理有数以千计的节点的图，不会像先前提到的GAN有低效率的问题。

文中关于Training主要是涉及了两个方面：

1. 采用Wasserstein GAN：为了提高稳定性。

2，为了控制生成的图与原来的图的相似程度，制定了两种早期停止策略：

1，VAL-CRITERION：训练过程中使用一个矩阵来作为过渡计数。这个矩阵的作用：评估验证集上的链路预测性能。停止方式：当性能改进停止时停止，也就是说在无法生成更加好的图的时候停止，这个完全时看机器自身的工作的。2，EO-CRITERION：达到符合用户期待的规格时就停止。更加的灵活，也使生成的图更加贴近原始的属性。这里就是看用户的意思来控制停止结束GAN训练的时间。

文章在实验部分将NetGAN拟合到CORA-ML和CITESER网络，以评估生成的图的质量。而经过各种比较作者认为除了NetGAN之外其余的生成图的模型相比于CORA-ML或多或少都会存在至少一个数据的偏移。作者还给出了一个显示了所有统计方法的平均排名，NetGAN的表现最好。其中EGRM似乎表现得出奇的好，但是它有严重的过度拟合，过度拟合说明实验与事实是不相符合的。相比之下，NetGAN在保存生成图中的属性以及泛化方面都做得很好。这里我认为作者的判断并没有说明的很清楚，他只是说出其余的模型都存在一定的数据的偏移，并没有说明这种偏移在何种程度内是符合要求的，也没有说明其余的模型是在哪个数据产生了怎样超出允许的偏移，因为经过细致的观察就可以发现NetGAN产生的数据也产生了偏移，而作者并有明确得说明这些偏移为何是符合实验目标的。

关于生成图的评价，作者给出了在不同的时间点的停止下图的同配性与训练的迭代的次数的关系以及边的重合与训练迭代的次数的关系：同配性随着迭代次数的增加而下降，而边的重合随时间数的增加而顺畅的增加。同配性和边的重合都是反应图的逻辑属性，这里也是说明了早期停止策略的在整个训练过程中的必要性。同时由于EO生成图的效果要好并且更加的灵活，作者更加倾向于这种停止策略。

之后为了找出相对最优的随机游走长度T，作者评估了在CORAML上改变T时链路预测性能的变化。首先训练具有不同随机游走长度的多个模型，并且评估确保每个模型观测到相同数目的过渡的分数。经验证实，该模型受益于使用较长的随机游走，而不是正好边缘（即T＝2）。t＝20在t＝16上的性能增益是边际的并且不超过额外的计算成本，简单地说就是在16,20上的选择上选择16，因为步长从16到20的优化上高额的计算成本与低效的优化不成正比。

作者在本文中主要着重于对单一复杂的大图的生成的处理，因为这也是NetGAN的优势所在：因为采用随机游走的方式，所以忽略了邻接矩阵中的0的元素，这样在处理大图时就可以比传统的方法更加有效率。但是，有关文章开头作者提出的问题，作者只是模糊地表示这种NetGAN因为采取了随机游走的方式，所以可以捕获图的拓扑性质，因此可以捕捉真实世界的图的潜在的未知属性。按照作者的想法，潜在的属性是可以通过随机游走的方式进行大量的训练获得的，也因此需要使用将0的元素排除这种方式否则就会效率很低。然而，对图的潜在属性的预测作者并没有通过真实的实验给出具体的数据，很多都只是对已有方法的一种优化。所以说，NetGAN的可行性理论上是可以的，但是实际的效果并没有体现出来。