DeepMerge II: 建立用于跨域并合星系识别的深度学习算法——复现报告

DeepMerge II: Build Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains

arxiv:2103.01373v1 DeepMergeDomainAdaptation

于浩然 2022.07.08

1 摘要

In astronomy, neural networks are often trained on simulation data with the prospect of being used on telescope observations. Unfortunately, training a model on simulation data and then applying it to instrument data leads to a substantial and potentially even detrimental decrease in model accuracy on the new target dataset. Simulated and instrument data represent different data domains, and for an algorithm to work in both, domain-invariant learning is necessary. Here we employ domain adaptation techniques—Maximum Mean Discrepancy (MMD) as an additional transfer loss and Domain Adversarial Neural Networks (DANNs)—and demonstrate their viability to extract domain-invariant features within the astronomical context of classifying merging and non-merging galaxies. Additionally, we explore the use of Fisher loss and entropy minimization to enforce better in-domain class discriminability. We show that the addition of each domain adaptation technique improves the performance of a classifier when compared to conventional deep learning algorithms. We demonstrate this on two examples: between two Illustris-1 simulated datasets of distant merging galaxies, and between Illustris-1 simulated data of nearby merging galaxies and observed data from the Sloan Digital Sky Survey. The use of domain adaptation techniques in our experiments leads to an increase of target domain classification accuracy of up to 20%. With further development, these techniques will allow astronomers to successfully implement neural network models trained on simulation data to efficiently detect and study astrophysical objects in current and future large-scale astronomical surveys.

在天文学中，经常在模拟数据上训练神经网络并可能将训练结果用于实际望远镜观测中。不幸的是，将在模拟数据上训练的模型应用于实际仪器数据将导致巨大且可能有害的在新的目标数据集上的准确度下降。模拟数据和观测数据代表不同的数据域，对于同时在两个域上工作的算法，域不变学习是必要的。这里我们

采用域适应技术：引入额外的迁移损失 (transfer loss)——最大平均偏差 (MMD)，以及域对抗神经网络 (DANNs)，并且证明了它们在识别并合星系的天文语境下导出域不变特征的可行性。此外我们探讨了 Fisher Loss 以及熵最小化 (Entropy Minimization) 来获得更好的域内类区分度 (in-domain class discriminabillity)。我们证明了，与传统的深度学习算法相比，每一项域适应技术的加入都提升了分类器的表现。我们用两个例子来阐述：两个 Illustris-1 模拟遥远星系的数据集；以及 Illustris-1 模拟的临近星系数据和斯隆数字巡天 (SDSS, 地面望远镜巡天项目) 观测到的实际临近星系数据集。我们实验中的域适应技术可使目标数据集的精度提升 20%。随着该技术进一步发展，这些技术将使天文学家成功使用在模拟数据上训练的神经网络模型有效地探测以及研究在现在和将来大规模巡天中的天体。

2 创新点

星系并合是宇宙早期星系演化的重要阶段，通常认为两个或多个旋涡星系并合会形成椭圆星系。研究并合星系有助于我们更好地理解星系中的恒星形成率、化学组成、粒子加速度以及其他性质；此外，它们对于宇宙学研究也相当重要，被用来研究宇宙中物质的演化。由于天文数据的庞大，常常使用机器学习方法来对各种天体进行分类。

然而，实际观测到的天文数据往往都没有标签 (unlabeled)。如果要对某一次的观测进行人工标记，想要取得数目可观的训练集需要耗费大量的精力和时间，而且不能保证标签的精度。于是，天文中训练神经网络模型通常使用计算机模拟的数据。但模拟数据和实际数据属于不同的域 (domains)，将在模拟数据上训练的模型直接用于实际数据得到的精度往往大约为 50%，和随机猜的精度相当，这样模型就失去意义了。于是，本文作者专门研究了各种域适应 (Domain Adaptation) 方法在并合星系识别下的效果，引入了 Fisher 损失、熵最小化与 MMD 方法和域对抗神经网络组合使用进行实验。

作者自己设计了一个名为 *DeepMerge* 的小型神经网络，由于其层数较少其更容易学习到域不变的特征，而不是纠结于对于特定域的细节特征。除了使用 *DeepMerge* 以外，作者还用 *Resnet18* 做了实验，得到了差不多的效果。这几种方法的引入使模型在训练集的精度提升了 20%，这对于天文领域是极大的进步。作者介绍了跨域学习的重要性，随着未来技术成熟有望用于更多的天文乃至自然科学场景。

3 代码结构

项目的主要文件结构如下：

主要程序文件为 `python_files` 路径下的 python 文件，其中，`no_domain_adaptation.py`，

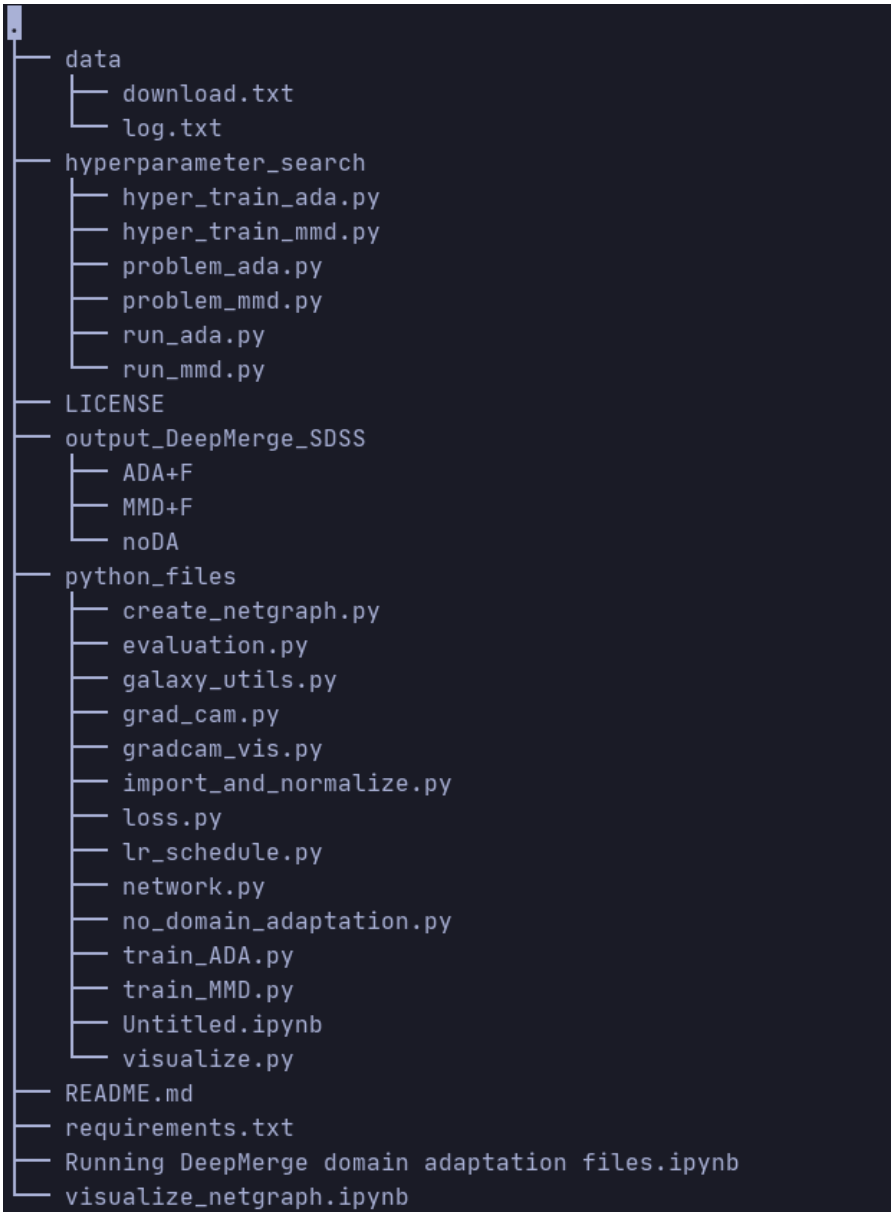


图 1: 项目文件结构树形图

`train_MMD.py` 以及 `train_ADA.py` 分别为按照不同方法进行训练的主程序,`network.py` 定义了训练所用的网络，其他是一些算法上的工具函数。源域和目标域的数据放在 `data` 目录下，用 wget 下载。`hyperparameter_search` 中是与超参数定义有关的函数。模型输出目录为 `output_DeepMerge_SDSS/`下的各个子目录，具体内容已省略。

4 训练过程

首先从 `requirements.txt` 安装依赖包。随后尝试用作者在 notebook 中的代码直接进行训练，发现 python 报错。看了源码，才发现作者提交的部分代码缩进为两格，费了一番功夫全部改规范，代码能够正常运行。

但由于第一次训练传参时忘记了传 `gpu_id`(实验室的 gpu 服务器是单卡的)，程序没能自动识别 gpu ，top 了一下看 cpu 占用率飙到 100%，而 nvidia-smi 显示显卡占用率为 0。后来由于没有将 cpu 训练保存的文件删除干净，后续改了参数也仍然用 cpu 训练，困惑了我一段时间。随后彻底清除生成文件后，能够正常使用 gpu 进行训练。

`tensorboard --log-dir output_DeepMerge_SDSS --bind-all`，在服务器上打开 tensorboard，在本地浏览器上可进行浏览。

文章中对无 DA、MMD、MMD+Fisher+Entropy、ADA、ADA+Fisher+Entropy 五种模型均在模拟-模拟集以及模拟-实际集上进行了训练，并在后面用 tSNE 方法将数据降维以可视化。在此复现报告中篇幅有限，难以将全部内容囊括，故只选取了模拟-实际数据集的 DA、MMD+Fisher+Entropy、ADA+Fisher+Entropy 三种模型进行训练。

这些模型的超参数经过了极其细微的优化调整，作者将他的超参给出如下图:

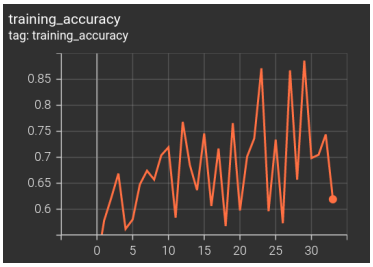
Table A3. The hyperparameters used to train DeepMerge on the simulated-to-real dataset. The second column of the table shows values for the baseline without domain adaptation, while the third column gives parameters for MMD and the fourth for MMD with transfer learning from the simulated-to-simulated dataset.

	DeepMerge: Simulated-to-Real		
Hyperparameters	noDA	MMD	TL+MMD
Learning rate	0.001	0.001	0.01
Beta	(0.7, 0.8)	(0.7, 0.8)	(0.7, 0.8)
Weight Decay	0.001	0.001	0.0001
Epsilon	10^{-8}	10^{-8}	10^{-8}
Cycle Length	5	5	5
Early stopping patience	20	20	20

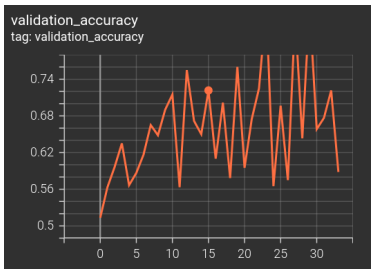
图 2: 超参数取值

当然这些超参也直接写在了 github 的程序里。按照表取超参，在终端采用传参的形式开始训练。在 tensorboard 中的截图展示如下:

4.1 noDA



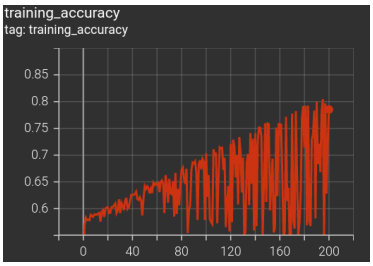
(a) 训练精度曲线



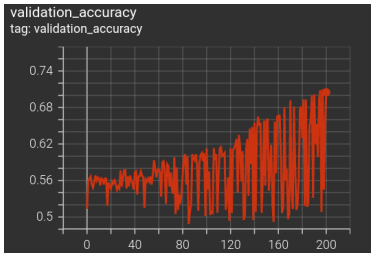
(b) 验证精度曲线

图 3: noDA(无域适应)

4.2 MMD+



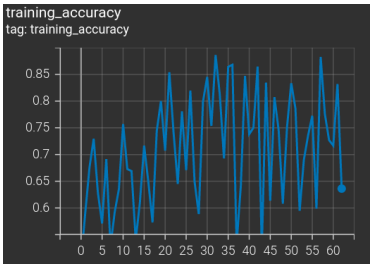
(a) 训练精度曲线



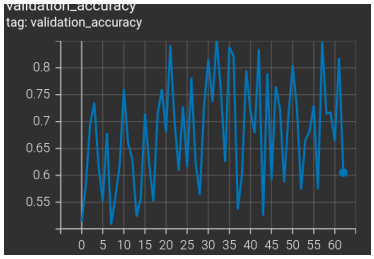
(b) 验证精度曲线

图 4: MMD+

4.3 ADA+



(a) 训练精度曲线



(b) 验证精度曲线

图 5: ADA+

4.4 讨论

明显可以看出 ADA+ 组的结果不理想，noDA 组的也类似，而 MMD+ 组的结果显得较为合理 (也不是那么合理)。在许多 epoch 以后，模型的精度仍有大量无规律的波动，这十分令人痛心。

5 复现结果与原文的不同

复现的结果并不理想，不能显著看出几种方法在目标域 (验证集) 上的差异。回顾原文献，发现作者采用了一定迁移学习 (Transfer Learning) 的方法，例如可以首先在 *imagenet* 这样不太相关的网络上先进行训练 (用和星系相关的数据训练效果会更好)，将预先训练好的模型作为检查点加载，可以得到性能的显著提升。

此外，作者采用了 *early stopping patience* 机制，当模型精度连续 20 epochs 没有变化时就会直接退出训练，而不是按照设定的数目训练 200 epoch。模型的效果不好也可能是由于 early stopping 时并没有达到真正的收敛。

此外，我十分怀疑作者的迁移学习部分是后来才加上的，因为如果不指定迁移学习所需 checkpoint 的路径，python 会直接报错退出。这是由于他使用了 `if config["ckpt_path"] is None:...`。当字典中不包含某一个键时，引用该键时会直接报错退出程序。我把这些代码改为 `if "ckpt_path" in config:...`，解决了这一问题。还有一种可能性，作者和我使用的 python 版本可能不同，因为他的 `requirements.txt` 以及 `README` 都没有提到他所使用的 python 版本。

综上，时间仓促，来不及把所有问题都修复好。炼丹不容易，能看出来作者得到这个好模型调参用了多大的精力，不过以后还是不太想再炼了。神经网络到底为什么能做出选择对人类来说还是为时尚早。