

机器学习纳米学位

猫狗大战——项目报告

2018 年 9 月 19 日

I. 问题的定义

项目概述

本项目意在构建一个深度学习模型，能够对任意给出的一张猫或狗的照片，判断究竟是猫还是狗。该问题涉及到计算机视觉这个领域，出发点是让计算机掌握识别图片的能力。

本项目将使用 **kaggle** 的猫狗项目的数据（包括已经标记好猫/狗标签的训练图片，和未做标记的测试图片），在计算机上对深度学习网络模型进行训练。

问题陈述

本项目的问题在于如何构建一个模型，用 **kaggle** 带标签的数据进行训练，然后对无标签的数据进行二分类预测（区分猫还是狗），这是一个监督学习的过程。我们可以对任意一张图片，预测该图为某个分类的概率，然后根据概率判定该图是否为某个分类。所以这个过程是可量化，可测量，可重复的。

整个问题的处理过程是，先用 **imagenet** 的预训练模型，对 **kaggle** 的训练、测试数据进行预测，导出特征向量；然后再构建自己的相对简单的 **MLP** 模型，以特征向量为输入，进行训练，得出二分类概率；最后用自己的 **MLP** 模型，对 **kaggle** 测试数据进行预测，看结果是否符合标准，并上传至 **kaggle** 进行打分。

评价指标

Logistic 回归损失函数，即对数损失函数，度量了真实条件概率分布与假定条件概率分布之间的差异，是常用的评价方式之一，被广泛应用于分类问题。而本项目涉及的是二分类问题，因此可以使用 **LogLoss** 作为评价标准。

评估标准采用 kaggle 官方指定损失函数：

$$LogLoss = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

参数说明

- n 为测试数据集的图片数量
- \hat{y}_i 为一张图片预测为狗的概率
- y_i 为类别标签，1 为狗，0 为猫
- $\log()$ 为自然对数

最终函数值越小，结果越好，代表模型性能越好。

项目要求是最低要达到 kaggle Public Leaderboard 前 10%，即排在第 $1314/10 = 131$ 名选手之前，其得分为 0.06127，即本项目最终得分要小于该数值。

II. 分析




(大概 2-4 页)

数据的探索

获取数据

在这一部分，你需要探索你将要使用的数据。数据可以是若干个数据集，或者输入数据/文件，甚至可以是一个设定环境。你需要详尽地描述数据的类型。如果可以的话，你需要展示数据的一些统计量和基本信息（例如输入的特征（features），输入里与定义相关的特性，或者环境的描述）。你还要说明数据中的任何需要被关注的异常或有趣的性质（例如需要做变换的特征，离群值等等）。你需要考虑：

从 kaggle 下载猫狗项目数据集，如果是在页面手动点击下载按钮，会得到一个 all.zip 文件，解压后会有 sample_submission.csv、test.zip、train.zip 共 3 个文件；如果是使用 kaggle API 命令（`kaggle competitions download -c dogs-vs-cats-redux-kernels-edition`）下载，会得到 test.zip、train.zip 共 2 个文件，须单独再下载 sample_submission.csv，最终数据的目录如下：

名称	大小
 sample_submission.csv	113.90 KB
 test.zip	284.48 MB
 train.zip	569.92 MB

解压 train.zip，里面包含 25000 张训练图片（猫狗各有 12500 张，文件名以 cat 或 dog 为前缀）；解压 test.zip，里面包含 12500 张测试图片（文件名为数字，无法区分猫狗）。

由于 keras 的 api 是通过不同目录来区分不同分类的，所以我们需要将 train 目录下的图片按 cat 或 dog 前缀划分到 2 个目录，test 目录下的图片再全部划分到一个 test 子目录中，最终目录如下：

--train

--cat

--cat.0.jpg

--cat.1.jpg

.....

--cat.12499.jpg

--dog

--dog.0.jpg

--dog.1.jpg

.....

--dog.12499.jpg

--test

--test

--1.jpg

--2.jpg

.....

--12500.jpg

随机展示几张图片，如下：



可以发现图片尺寸不一。

检测异常数据

本项目使用 `keras` 预训练模型 `ResNet50` (`imagenet` 权重)，对训练数据集的图片进行预测（会给出图片分别属于不同种类的概率），筛选出 `top100` 概率都不是狗/猫种类的图片，然后人工判断是否合理，最后选择其中的真正异常图片进行剔除。步骤如下：

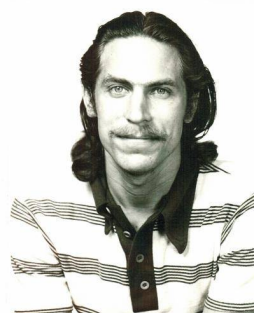
1、先对部分图片进行检测，刚开始使用 `top5`，发现结果里面正常图片占比较大，则不断调整 `top` 的种类数量，当为 `100` 时，发现异常图片的漏测率、误测率都较低，就选了这个参数，

然后对所有训练图片进行检测：

对`./data/train/dog`目录下的图片进行预测，筛选出 `top100` 概率都不是狗的图片，打印其文件名称。

对`./data/train/cat`目录下的图片进行预测，筛选出 `top100` 概率都不是猫的图片，打印其文件名称。

2、对这些图片进行人工判断，筛选出异常图片，移动到 `invalid_train` 相应子目录下，随机展示几张异常图片（包括图片本身就不是猫狗，或者是图画，或者是背景太复杂的情况）：



探索性可视化

在这一部分，你需要对数据的特征或特性进行概括性或提取性的可视化。这个可视化的过程应该要适应你所使用的数据。就你为何使用这个形式的可视化，以及这个可视化过程为什么是有意义的，进行一定的讨论。你需要考虑的问题：

- 你是否对数据中与问题有关的特性进行了可视化？
- 你对可视化结果进行详尽的分析和讨论了吗？
- 绘图的坐标轴，标题，基准面是不是清晰定义了？

算法和技术

在这一部分，你需要讨论你解决问题时用到的算法和技术。你需要根据问题的特性和所属领域来论述使用这些方法的合理性。你需要考虑：

- 你所使用的算法，包括用到的变量/参数都清晰地说明了吗？
- 你是否已经详尽地描述并讨论了使用这些技术的合理性？
- 你是否清晰地描述了这些算法和技术具体会如何处理这些数据？

基准模型

在这一部分，你需要提供一个可以用于衡量解决方案性能的基准结果/阈值。这个基准模型要能够和你的解决方案的性能进行比较。你也应该讨论你为什么使用这个基准模型。一些需要考虑的问题：

- 你是否提供了作为基准的结果或数值，它们能够衡量模型的性能吗？
- 该基准是如何得到的（是靠数据还是假设）？

III. 方法

(大概 3-5 页)

数据预处理

在这一部分，你需要清晰记录你所有必要的的数据预处理步骤。在前一个部分所描述的数据的异常或特性在这一部分需要被更正和处理。需要考虑的问题有：

- 如果你选择的算法需要进行特征选取或特征变换，你对此进行记录和描述了吗？

- *数据的探索这一部分中提及的异常和特性是否被更正了，对此进行记录和描述了吗？*
- *如果你认为不需要进行预处理，你解释个中原因了吗？*

执行过程

在这一部分，你需要描述你所建立的模型在给定数据上执行过程。模型的执行过程，以及过程中遇到的困难的描述应该清晰明了地记录和描述。需要考虑的问题：

- *你所用到的算法和技术执行的方式是否清晰记录了？*
- *在运用上面所提及的技术及指标的执行过程中是否遇到了困难，是否需要作出改动来得到想要的结果？*
- *是否有需要记录解释的代码片段(例如复杂的函数)？*

完善

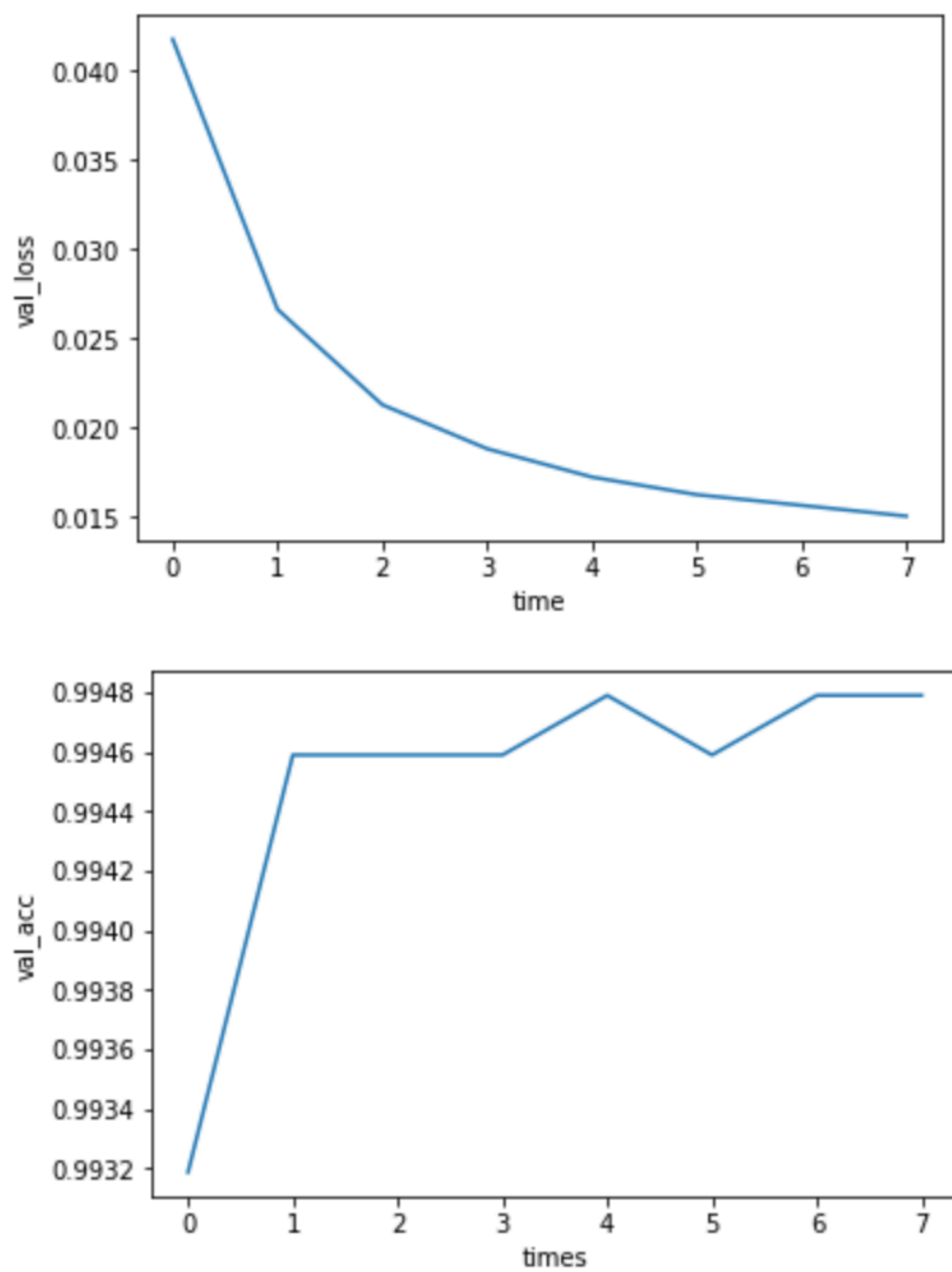
在这一部分，你需要描述你对原有的算法和技术完善的过程。例如调整模型的参数以达到更好的结果的过程应该有所记录。你需要记录最初和最终的模型，以及过程中有代表性意义的结果。你需要考虑的问题：

- *初始结果是否清晰记录了？*
- *完善的过程是否清晰记录了，其中使用了什么技术？*
- *完善过程中的结果以及最终结果是否清晰记录了？*

IV. 结果


(大概 2-3 页)

模型的评价与验证



上图为模型的学习曲线及准确率。

最终 kaggle 得分 0.04156，小于 0.06127，达到了项目要求：



Dogs vs. Cats Redux: Kernels Edition

Distinguish images of dogs from cats
1,314 teams · 2 years ago

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Late Submission](#)

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
predict.csv	2 minutes ago	0 seconds	0 seconds	0.04156

Complete

[Jump to your position on the leaderboard ▾](#)

合理性分析

模型最终收敛，得分比基准的效果好，达到了较高准确率，因此模型是有效的，也解决了项目中设定的问题。

V. 项目结论

(大概 1-2 页)

结果可视化

在这一部分，你需要用可视化的方式展示项目中需要强调的重要技术特性。至于什么形式，你可以自由把握，但需要表达出一个关于这个项目重要的结论和特点，并对此作出讨论。一些需要考虑的：

- 你是否对一个与问题，数据集，输入数据，或结果相关的，重要的技术特性进行了可视化？
- 可视化结果是否详尽的分析讨论了？
- 绘图的坐标轴，标题，基准面是不是清晰定义了？

对项目的思考

在这一部分，你需要从头到尾总结一下整个问题的解决方案，讨论其中你认为有趣或困难的地方。从整体来反思一下整个项目，确保自己对整个流程是明确掌握的。需要考虑：

- 你是否详尽总结了项目的整个流程？
- 项目里有哪些比较有意思的地方？
- 项目里有哪些比较困难的地方？
- 最终模型和结果是否符合你对这个问题的期望？它可以在通用的场景下解决这些类型的问题吗？

需要作出的改进

在这一部分，你需要讨论你可以怎么样去完善你执行流程中的某一方面。例如考虑一下你的操作的方法是否可以进一步推广，泛化，有没有需要作出变更的地方。你并不需要确实作出这些改进，不过你应能够讨论这些改进可能对结果的影响，并与现有结果进行比较。一些需要考虑的问题：

- 是否可以有算法和技术层面的进一步的完善？
- 是否有一些你了解到，但是你还没能够实践的算法和技术？
- 如果将你最终模型作为新的基准，你认为还能有更好的解决方案吗？

**** 在提交之前，问一下自己... ****

- 你所写的项目报告结构对比于这个模板而言足够清晰了没有？
- 每一个部分（尤其**分析**和**方法**）是否清晰，简洁，明了？有没有存在歧义的术语和用语需要进一步说明的？
- 你的目标读者是不是能够明白你的分析，方法和结果？
- 报告里面是否有语法错误或拼写错误？
- 报告里提到的一些外部资料及来源是不是都正确引述或引用了？
- 代码可读性是否良好？必要的注释是否加上了？
- 代码是否可以顺利运行并重现跟报告相似的结果？

参考文献