机器学习工程师

毕业项目开题报告

沈捷

2019年3月28日

# 背景

2013年，kaggle举办了一项趣味性赛事“猫狗大战”，即猫狗图像分类。许多学者及行业人士参与其中，供献了大量的解决方案，机器学习领域的局势由之改变。现今图像分类在许多行业已展开深入研究，如医学上热门的癌组织与良性组织的鉴别，甚至有学者尝试利用深度学习方法鉴别出男女大脑MRI图像的差异<cite data-cite="2636952/IJ2W7TIB"></cite>，都是类似的问题。但医学领域常由于患者隐私保护政策，难以获得大量的训练数据，所以为了更好地学习、研究新的方法，也有学者会利用猫狗大战数据进行探索尝试和验证<cite data-cite="2636952/YVAJHILJ"></cite>，同理可以延伸至许多应用领域。

### 问题描述

区分图片是猫还是狗的照片，是一个明显的二分类问题，通过输入图像特征，获得一个概率，通过概率来判断属于哪一个类。

### 数据输入

该数据集已切分为训练集和测试集。训练集包含25000张图片，猫和狗图片比例为1:1，可以从文件名（\*标签.编号.jpg\*）中获得标签；测试集有12500张图片，文件名只有编号。

训练集图片会经过随机旋转、缩放、裁剪等预处理，使模型能有更好的泛化能力。这些图像将统一为$hh \times ww$大小，并分解为RGB三个颜色通道的色值，再经过标准化处理，作为模型的输入。

### 解决方案

\_(approx. 1 paragraph)\_

处理好输入数据后，将采用一个预训练的神经网络进行迁移学习，预计将采用vgg16模型。神经网络的输入节点应为$hh \times ww$大小，输出为1，采用sigmoid激活函数获得图片指向某一类的概率（狗=1，猫=0）。最后将模型在测试集上进行预测，提交到kaggle上进行比较判断。模型的训练目标即是达到kaggle的前10%。

训练模型的过程中，将调整epoch、隐藏层节点、学习速率等超参数，以尽量高准确率，降低Loss.

附：sigmoid函数公式

### 基准模型

\_(approximately 1-2 paragraphs)\_

没有基准模型。

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.

### 评估指标

\_(approx. 1-2 paragraphs)\_

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

### 项目设计

\_(approx. 1 page)\_

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

### 参考文献

<div class="cite2c-biblio"></div>

-----------

\*\*Before submitting your proposal, ask yourself. . .\*\*

- Does the proposal you have written follow a well-organized structure similar to that of the project template?

- Is each section (particularly \*\*Solution Statement\*\* and \*\*Project Design\*\*) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?

- Would the intended audience of your project be able to understand your proposal?

- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?

- Are all the resources used for this project correctly cited and referenced?