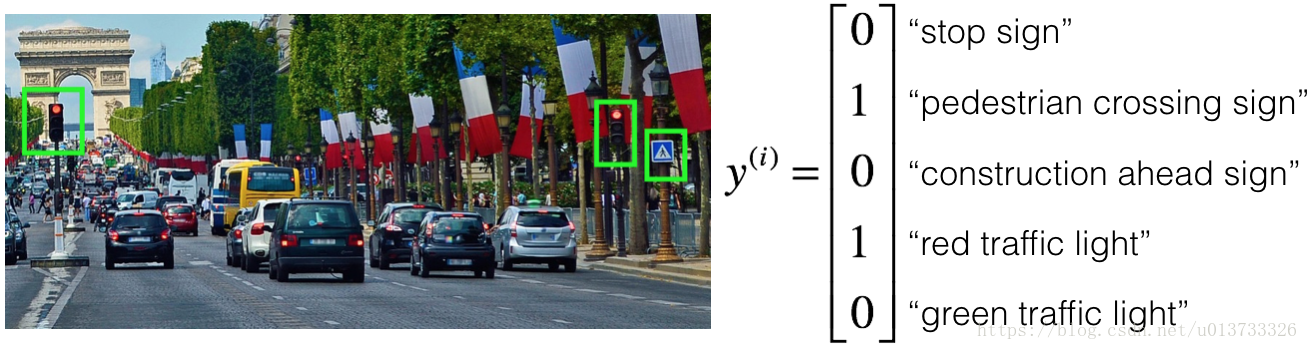
# 自动驾驶(案例研究)

**1、**为了帮助练习机器学习策略，本周我们将介绍另一种场景并询问你将如何做。我们认为这个在机器学习项目中工作的“模拟器”将给出一个引导机器学习项目的任务。



你受雇于一家创业的自动驾驶的创业公司，负责检测图片中的路标（停车标志，行人过路标志，前方施工标志）和交通信号标志（红灯和绿灯），目标是识别哪些对象出现在每个图片中。例如，上面的图片包含一个行人过路标志和红色交通信号灯标志。   
  您的100,000张带标签的图片是使用您汽车的前置摄像头拍摄的，这也是你最关心的数据分布，您认为您可以从互联网上获得更大的数据集，即使互联网数据的分布不相同，这也可能对训练有所帮助。**你刚刚开始着手这个项目，你做的第一件事是什么？**假设下面的每个步骤将花费大约相等的时间（大约几天）。

【 】 花几天时间去获取互联网的数据，这样你就能更好地了解哪些数据是可用的。

【 】 花几天的时间检查这些任务的人类表现，以便能够得到贝叶斯误差的准确估计。

【 】 花几天的时间使用汽车前置摄像头采集更多数据，以更好地了解每单位时间可收集多少数据。

【★】 花几天时间训练一个基本模型，看看它会犯什么错误。(正如在视频中多次看到的，机器学习是一个高度迭代的过程。我们需要在基本模型上创建、编码和实验，然后迭代以找出对给定问题最有效的模型。)

**2、**您的目标是检测道路标志（停车标志、行人过路标志、前方施工标志）和交通信号（红灯和绿灯）的图片，目标是识别这些图片中的哪一个标志出现在每个图片中。 您计划在隐藏层中使用带有ReLU单位的深层神经网络。

**对于输出层，使用Softmax激活将是输出层的一个比较好的选择，因为这是一个多任务学习问题，对吗？**

【 】 True

【★】 False (如果每个图片中只有一个可能性：停止标志、减速带、人行横道、红绿灯， 那么SoftMax将是一个很好的选择。由于不是这种情况，所以不能使用Softmax 激活函数。)

**3、你正在做误差分析并计算错误率，在这些数据集中，你认为你应该手动仔细地检查哪些图片（每张图片都做检查）？**

【 】 随机选择10,000图片

【 】 随机选择500图片

【★】 500张算法分类错误的图片。(查看算法分类出错的那些图片是非常重要的，由于查看算法分类错误造成的每个图片都不太实际，所以我们需要随机选择500个这样的图片并分析出现这种错误的原因。)

【 】 10,000张算法分类错误的图片。

**4、**在处理了数据几周后，你的团队得到以下数据：

- 100,000张使用汽车前摄像头拍摄的标记了的图片。

- 900,000张从互联网下载的标记了道路的图片。

每张图片的标签都精确地表示任何的特定路标和交通信号的组合。例如，表示图片包含了停车标志和红色交通信号灯。

因为这是一个多任务学习问题，你需要让所有向量被完全标记。 **如果一个样本等于，那么学习算法将无法使用该样本，是正确的吗？**

【 】 正确

【★】 错误(在多任务学习的视频中，您已经看到，即使某些条目没有被标记，您也可以计算成本。该算法不会受到数据中某些条目未标记的样本的影响。)

**5、你所关心的数据的分布包含了你汽车的前置摄像头的图片，这与你在网上找到并下载的图片不同。如何将数据集分割为训练/开发/测试集?**

【 】 将10万张前摄像头的图片与在网上找到的90万张图片随机混合，使得所有数据都随机分布。 将有100万张图片的数据集分割为：有60万张图片的训练集、有20万张图片的开发集和有20万张图片的测试集。

【 】 将10万张前摄像头的图片与在网上找到的90万张图片随机混合，使得所有数据都随机分布。将有100万张图片的数据集分割为：有98万张图片的训练集、有1万张图片的开发集和有1万张图片的测试集。

【★】 选择从互联网上的90万张图片和汽车前置摄像头的8万张图片作为训练集，剩余的2万张图片在开发集和测试集中平均分配。(正如在课堂上看到的那样，分配数据的方式非常重要，您的训练和开发集的分布类似于“现实生活”数据。此外，测试集应包含您实际关心的足够数量的“现实生活”数据。)

【 】 选择从互联网上的90万张图片和汽车前置摄像头的2万张图片作为训练集，剩余的8万张图片在开发集和测试集中平均分配。  
**6、**假设您最终选择了以下拆分数据集的方式:



**您还知道道路标志和交通信号分类的人为错误率大约为0.5％。以下哪项是真的（检查所有选项）?**

【 】 由于开发集和测试集的错误率非常接近，所以你过拟合了开发集。

【★】 你有一个很大的数据不匹配问题，因为你的模型在训练-开发集上比在开发集上做得好得多。

【★】 你有一个很大的可避免偏差问题，因为你的训练集上的错误率比人为错误率高很多。

【 】你有很大的方差的问题，因为你的训练集上的错误率比人为错误率要高得多。

【 】 你有很大的方差的问题，因为你的模型不能很好地适应来自同一训练集上的分布的数据，即使是它从来没有见过的数据。

**7、根据上一个问题的表格，一位朋友认为训练数据分布比开发/测试分布要容易得多。你怎么看？**

【 】 你的朋友是对的。（即训练数据分布的贝叶斯误差可能低于开发/测试分布）。

【 】 你的朋友错了。（即训练数据分布的贝叶斯误差可能比开发/测试分布更高）。

【★】 没有足够的信息来判断你的朋友是对还是错。(为了了解这一点，我们必须在两个分布上分别测量人的水平误差，该算法对训练的分布数据有更好的效果。但我们不确定这是因为它被训练在数据上，或者它比开发/测试分布更容易。)

【 】 无论你的朋友是对还是错，这些信息都对你没有用。

**8、**您决定将重点放在开发集上, 并手动检查是什么原因导致的错误。下面是一个表, 总结了您的发现:



在这个表格中，4.1％、8.0％这些比例是总开发集的一小部分（不仅仅是您的算法错误标记的样本），即大约8.0 / 14.3 = 56％的错误是由于雾天的图片造成的。

**从这个分析的结果意味着团队最先做的应该是把更多雾天的图片纳入训练集，以便解决该类别中的8%的错误，对吗？**

【★】 错误，因为这取决于添加这些数据的容易程度以及您要考虑团队认为它会有多大帮助。

【 】 是的，因为它是错误率最大的类别。正如视频中所讨论的，我们应该对错误率进行按大小排序，以避免浪费团队的时间。

【 】 是的，因为它比其他的错误类别错误率加在一起都大(8.0 > 4.1 + 2.2 + 1.0)。

【 】 错误，因为数据增强(通过清晰的图像+雾的效果合成雾天的图像)更有效。

**9、你可以买一个专门设计的雨刮，帮助擦掉正面相机上的一些雨滴。根据上一个问题的表格，您同意以下哪些陈述？**

【★】 对于挡风玻璃雨刷可以改善模型的性能而言，2.2％是改善的最大值。(一般而言，解决了雨滴的问题你的错误率可能不会完全降低2.2%，如果你的数据集是无限大的, 改善2.2% 将是一个理想的估计, 买一个雨刮是应该可以改善性能的。)

【 】对于挡风玻璃雨刷可以改善模型的性能而言，2.2％是改善最小值。

【 】对于挡风玻璃雨刷可以改善模型的性能而言，改善的性能就是2.2％。

【 】在最坏的情况下，2.2%将是一个合理的估计，因为挡风玻璃刮水器会损坏模型的性能。

**10、**您决定使用数据增强来解决雾天的图像，您可以在互联网上找到1,000张雾的照片，然后拿清晰的图片和雾来合成雾天图片，如下所示：



**你同意下列哪种说法？（检查所有选项）**

【 】 只要你把它与一个更大（远大于1000）的清晰/不模糊的图像结合在一起，那么对雾的1000幅图片就没有太大的过拟合的风险。

【 】 将合成的看起来像真正的雾天图片添加到从你的汽车前摄像头拍摄到的图片的数据集对与改进模型不会有任何帮助，因为它会引入可避免的偏差。

【★】 只要合成的雾对人眼来说是真实的，你就可以确信合成的数据和真实的雾天图像差不多，因为人类的视觉对于你正在解决的问题是非常准确的。(如果合成的图像看起来逼真, 就好像您在有雾的天气中添加了有用的数据来识别道路标志和交通信号一样。)

**11、在进一步处理问题之后，您已决定更正开发集上错误标记的数据。您同意以下哪些陈述？（检查所有选项）。**

【★】 您不应更正训练集中的错误标记的数据, 以免现在的训练集与开发集更不同。 (深度学习算法对于略有不同的训练集和开发集分布是相当强大的。（博主注：意思是小改动会造成大差异）)

【 】 您应该更正训练集中的错误标记数据, 以免您现在的训练集与开发集更不同。

【 】 您不应该更正测试集中错误标记的数据，以便开发和测试集来自同一分布。

【★】 您还应该更正测试集中错误标记的数据，以便开发和测试集来自同一分布。(因为你想确保你的开发和测试数据来自相同的分布，以使你的团队的迭代开发过程高效。)

**12、**到目前为止，您的算法仅能识别红色和绿色交通灯，该公司的一位同事开始着手识别黄色交通灯（一些国家称之为橙色光而不是黄色光，我们将使用美国的黄色标准），含有黄色灯的图像非常罕见，而且她没有足够的数据来建立一个好的模型，她希望你能用转移学习帮助她。

**你告诉你的同事怎么做？**

【★】 她应该尝试使用在你的数据集上预先训练过的权重，并用黄光数据集进行进一步的微调。(你已经在一个庞大的数据集上训练了你的模型，并且她有一个小数据集。 尽管您的标签不同，但您的模型参数已经过训练，可以识别道路和交通图像的许多特征，这些特征对于她的问题很有用。 这对于转移学习来说是一个完美的例子，她可以从一个与您的架构相同的模型开始，改变最后一个隐藏层之后的内容，并使用您的训练参数对其进行初始化。)

【 】 如果她有10,000个黄光图像，从您的数据集中随机抽取10,000张图像，并将您和她的数据放在一起，这可以防止您的数据集“淹没”她的黄灯数据集。

【 】 你没办法帮助她，因为你的数据分布与她的不同，而且缺乏黄灯标签的数据。

【 】 建议她尝试多任务学习，而不是使用所有数据进行迁移学习。

**13、**另一位同事想要使用放置在车外的麦克风来更好地听清你周围是否有其他车辆。例如，如果你身后有警车，你就可以听到警笛声。**但是，他们没有太多的训练这个音频系统，你能帮忙吗？**

【 】从视觉数据集迁移学习可以帮助您的同事加快步伐，多任务学习似乎不太有希望。

【 】 从您的视觉数据集中进行多任务学习可以帮助您的同事加快步伐，迁移学习似乎不太有希望。

【 】迁移学习或多任务学习可以帮助我们的同事加快步伐。

【★】迁移学习和多任务学习都不是很有希望。(他试图解决的问题与你的问题完全不同，不同的数据集结构可能无法使用迁移学习或多任务学习。)

**14、**要识别红色和绿色的灯光，你一直在使用这种方法：

A：将图像*x*输入到神经网络，并直接学习映射以预测是否存在红光(和/或)绿光*y*。

一个队友提出了另一种两步走的方法：

B：在这个两步法中，您首先要检测图像中的交通灯（如果有），然后确定交通信号灯中照明灯的颜色。

**在这两者之间，方法B更多的是端到端的方法，因为它在输入端和输出端有不同的步骤，这种说法正确吗？**

【 】 正确

【★】 错误 (A是一种端到端的方法，因为它直接将输入（x）映射到输出（y）。)

**15、如果你有一个什么，在上面的问题中方法A往往比B方法更有效**

【★】 大训练集 (在许多领域，据观察，端到端学习在实践中效果更好，但需要大量数据。 如果没有大量的数据，端到端深度学习的应用是效果比较差的。)

【 】 多任务学习的问题。

【 】 偏差比较大的问题。

【 】 高贝叶斯误差的问题。