Don’t take the easy way out ：Ensemble based methods for avoiding known dataset biases论文阅读

<https://blog.csdn.net/bxg1065283526/article/details/106163357?utm_medium=distribute.pc_relevant.none-task-blog-OPENSEARCH-4.control&depth_1-utm_source=distribute.pc_relevant.none-task-blog-OPENSEARCH-4.control>

1. Abstract

最先进的模型通常会利用数据间的表层形式，这些模式通常无法很好的推广到数据集外或对抗性环境中。例如：文本模型通常会学习到的特定的关键词隐含的含义，而视觉问答模型则学习预测原型答案，而不考虑图像中的证据。

在本文中，作者证明，如果事先知道数据集偏差，则作者可以训练一个模型，使其对域移位更具健壮性。

作者方法分为两个阶段

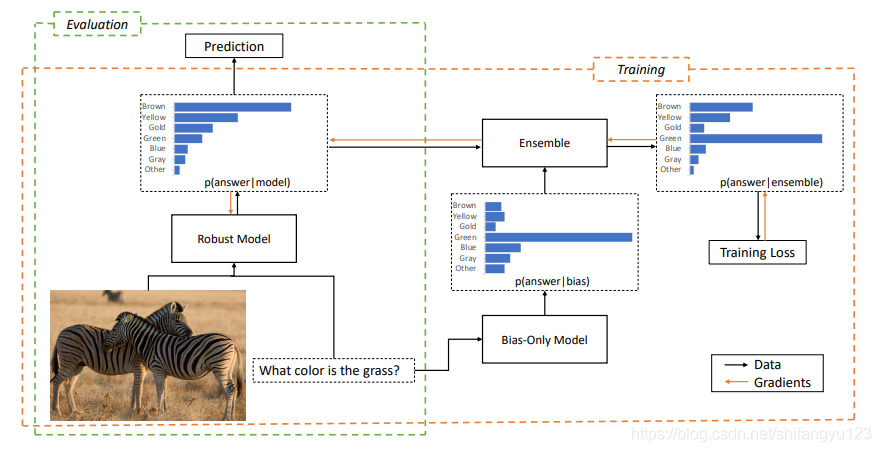
* 训练一个纯粹基于数据集偏差进行预测的朴素模型
* 训练一个健壮的模型，作为与朴素模型集成的一部分，以鼓励它关注数据中更可能泛化的其他模式。

1. Introduction

在本文中，作者在这些工作的基础上展示，一旦数据集偏差被识别，我们可以通过防止模型利用该偏差来提高模型的域外性能。为了做到这一点，我们利用了这样一个事实，即这些偏差通常是可以用简单的，受约束的基线方法建模，以便通过基于集成的训练将他们从最终模型中剔除。

作者的方法分为两步，第一步训练一个带有偏见的模型，模型在训练集表现好，但是在这范围之外表现就很差，第二步再训练一个模型集成带偏见的模型，在测试集上只用第二个模型。

作者首先在两个通过向MNLI添加人工特征来包含人工构建的偏差的合成数据集上验证，然后再在三个有挑战性的数据集（这里包含VQA-CP），作者在所有设定上都取得了很大的进步。



1. Method

这一部分描述了方法的两个阶段，

* 建立一个仅问题的模型，
* 使用它融合训练一个健壮的模型
  1. 训练仅问题模型

第一阶段的目标构建一个在训练数据上表现良好的模型，但是在域外测试集上可能表现很差。因为我们假设我们无法访问测试集中的示例，所以我们必须应用先验知识来实现该目标，最直接的方法就是识别一组特征，这些特征在训练期间与类标签相关，但已知与测试集上的标签不相关或反相关，然后根据这些特征训练分类器。

* 1. 训练一个健壮的模型

这一阶段训练出一个健壮模型，避免使用仅偏差模型学习的方法。

* + 1. 问题定义

我们假设n个训练例子<x1、x2、x3…xn>，每个都有一个整数标记yi，其中yi属于{1,2,3,…,C}C为类的数目。

另外，我们假设一个预训练的偏差预测变量h，其中h（xi）=bi=<bi1、bi2、…、bic>，bij是仅偏差模型对样本i是j类的预测概率。

最后，我们有第二个预测函数f，参数为θ，其中

*f*（*xi*,*θ*） = pi

是类的相似概率分布

我们的目标是构建一个优化θ的训练目标，以便f可以学习选择正确的类别，而无需使用偏差模型所捕获的策略。

* + 1. Learned-Mixin

我们训练h和f的集合，对每个示例，通过组合pi和bi来计算新的类别分布。

在训练期间，使用计算损耗，并通过f反向传播梯度，在评估期间单独使用f。我们提出几种不同的组合方法。

我们允许在模型明确在给定输入的情况下信任偏差的程度：

*p̂i =softmax*（*log*（*pi*）*+g*（*xi*）*log*（*bi*））

其中g是学习函数，我们将g计算为

其中w是学习向量

hi是模型的最后一个隐藏层

例如xi，并且函数用于防止模型通过将负数乘以负权值来反转其偏向。 用其余的模型参数训练w。 当时，这减少了偏差乘积。这种方法的一个难点是模型可以学习将偏差积分到pi中并将g（xi）设置为0。

* + 1. Learned-Mixin+H

为了防止learned-mixin集合忽略bi，我们对损失添加了熵罚：

*R = wH*( *softmax*( *g*( *xi* )*log*( *bi* )))

其中H（z）是熵，w是超参数，惩罚熵会鼓励偏差分量不均匀，从而对整体产生更大的影响。

1. Experiment

