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

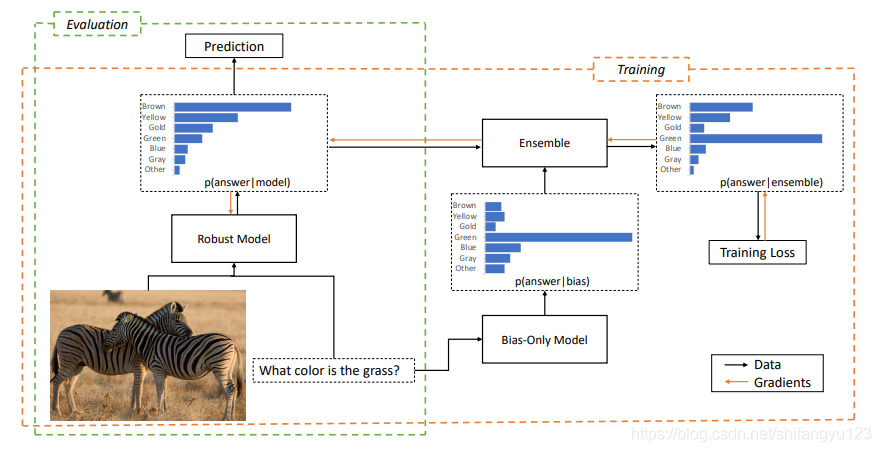
* 训练一个纯粹基于数据96C6偏差进行预测的朴素模型
* 训练一个健壮的模型，作为与朴素模型集成的一部分，以鼓励它关注数据中更可能泛化的其他模式。

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. The Related Work

之前提出的数据集很多数据集中都存在偏见，新提出的数据集在尽量避免偏见，作者认为虽然避免数据集存在偏见很重要，但提出一个方法能够防止模型利用之前的数据集中的偏见，然后之前的数据集就可以继续使用，随着我们对我们想要避免的偏见的理解的发展，更新我们的方法。

之前关于偏见（忽略输入的一部分而产生的）的工作主要集中在强制生成模型以理解所有输入的生成目标，精心设计的模型体系结构或从模型的内部表示对抗性删除类指示特征。

作者考虑除此之外的偏见，并在VQA-CP数据集上验证。还有一个相关的任务是防止模型使用特定的有问题的数据集特征，这通常是从公平角度进行研究的。但作者任务偏见是跟整个任务的特征有关，不能简单的忽略。通过对现有示例应用较小扰动而构建的域外示例的模型评估也成为近期研究的主题。 作者考虑的领域转移涉及对输入分布的较大更改，旨在消除现有模型中的更高级别的缺陷。

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。

最直接的方法就是在训练过程中识别与类别标签相关但已知与测试集上标签不相关的一组特征，然后在这些特征上训练分类器。例如，我们的VQA-CP纯偏差模型使用问题类型作为输入，因为问题类型和答案之间的相关性在训练集中与测试集有很大不同（2是“多少个。 。”这类问题的常见答案，但对于测试装置上的此类问题很少见。）

但是，我们方法的好处是可以使用任何一种预测变量对偏差进行建模，从而为我们提供了一种捕获更复杂直觉的方法。例如，在SQuAD上，我们的“仅偏见”模型会根据TF-IDF分数构建的输入进行操作，而在我们更改的先前TriviaQA数据集上，我们的“仅偏见”模型会使用预先训练的命名实体识别（NER）标记器 。

* 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个训练样本<x1、x2、x3…xn>，每个都有一个整数标签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器（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，我们有第二个预测器（函数）f，参数为θ，其中

*f*（*xi*,*θ*） = pi

是类的相似概率分布

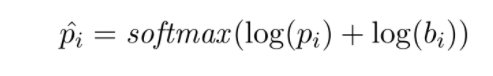
Pi也是一个类似的类别概率。

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. General Approach

我们将h和f进行融合，特别是一个新类别的概率的分布p̂i是由pi和bi计算得出。在训练期间，loss的计算是通过p̂i，梯度反向传播f，在验证阶段，只有*f*使用。

* + 1. Bias product

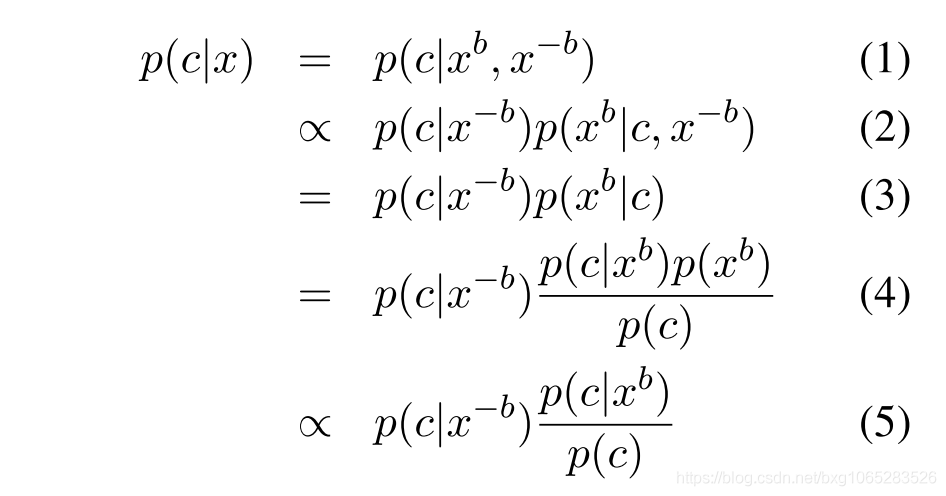


* 无限接近于，代表两个数组进行element-wise相乘

**Probabilistic Justificaition**

给定样本x，xb代表样本的bias，在只有bias的模型中用到。x-b代表样本中可被捕捉的除bias之外的信息。

假定给定标签c，xb和x-b是有条件独立的。



我们无法直接对p(c|x^{-b})建模，因为通常无法创建排除偏差（bias）的数据视图。相反，为了鼓励模型落入计算p(c|x^{-b})的角色，我们使用仅偏差模型计算p(c | x^b)/p(c)，两个模型的乘积来计算p(c|x)。

实际上，我们忽略了p（c）因子，因此，在我们的数据集上，要么类是均匀分布（MNLI）的，要么纯偏差模型无法轻松捕获类，因为它正在使用指针网络（QA）；或者我们还是想从模型（VQA）中删除类优先级。

* + 1. Learned-Mixin

条件独立性（等式3）的假设通常太强了，例如，在某些情况下，健壮模型可能能够预测仅偏倚模型对于某些种类的训练示例而言将是不可靠的。我们发现这可能会导致健壮模型选择性调整其行为，已补偿偏差模型的不准确性，从而导致域外设置出现错误。（5.1）

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g=softolus(w \cdot h_i)

其中w是学习向量

hi是样本xi的模型的最后一层

softplus(x)=log(1+e^x)，用于通过将模型乘以负权重来防止模型逆转偏差。W与其他参数一起训练，，当g(x_i)= 1时，这减少了偏差乘积。

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。

这种方法的一个困难之处就是可以学习将偏差积分到*pi*中并将g(x_i)设置为0，我们发现这种情况有时在实践中会发生，我们的下一种方法可以缓解这一挑战。

* + 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。

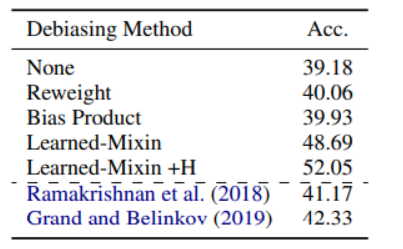
1. Experiment

在几个具有域外测试集的数据集上评估我们的方法。其中一些任务，如HANS或Adversarial SQuAD，可以很容易地通过生成与测试集中类似的额外训练示例来解决。相反，我们证明了通过利用模型可能采用的一般的、有偏见的策略的知识来提高这些任务的性能是可能的。

我们的评估设置包括训练集，域外测试集，纯偏差模型和主模型。 为了进行评估，我们在训练集上训练纯偏差模型，在使用第3节中的一种方法的同时，在训练集上训练主模型，并在域外测试集上评估主模型。 如果可用，我们还会报告域内测试集的性能。 我们使用的模型在主模型中可以很好地完成其各自的任务，并且不会进一步调整其超参数或进行早期停止。

我们考虑了两个提取的质量检查数据集，我们将其视为联合分类任务，其中模型必须选择开始和结束答案令牌。 对于这些数据集，我们建立了独立的仅偏倚模型来选择开始标记和结束标记，然后将这些偏见与分类器的开始标记和结束标记的输出分布结合在一起。 我们在问题和段落嵌入中应用ReLU层，然后进行最大池化，以构造一个隐藏状态来计算学习的mixin权重。

我们将我们的方法与以下所述的重新加权基准进行比较，并与未经任何修改地训练主模型进行了比较。 在VQA上，我们还与对抗方法进行了比较。 我们考虑的其他偏差并非仅基于观察部分输入，因此这些对抗方法无法直接应用。



* 1. Rewight Baseline

作为非整体基线，我们在数据的加权版本上训练主模型，其中示例x_i的权重为1-b_{iy_i}(即，我们将示例权重为1减去仅偏向模型分配正确标签的概率 ）。 这鼓励主模型专注于仅偏见模型出错的示例。

* 1. Hyperparameters

我们的一种方法（Learned-Mixin + H）需要超参数调整。 但是，由于我们的假设是在培训期间我们无法访问域外测试示例，因此在我们的设置中，超参数调整面临挑战。 一个合理的选择是调整展示给测试集相关但不相同的域转移的开发集上的超参数，但不幸的是，我们的数据集中都没有这样的开发集。 相反，我们遵循先前的工作（Grand and Belinkov，2019; Ramakrishnan et al。，2018），并在测试集上执行模型选择。 尽管这对该方法的结果提出了重要的警告，但我们认为仍然有必要观察到熵正则化器可以产生很大的影响。 未来的工作可能能够构建合适的开发集，或者提出其他超参数调整方法来缓解此问题。 所选的超参数显示在附录A中。

1. Experiments

因为我只对VQA任务感兴趣，所以只列一下VQA-CP

* 1. VQA-CP

**Data:**我们对VQA-CP v2（Agrawal等人，2018）数据集进行了评估，该数据集是通过将VQA 2.0（Goyal等人，2018）训练和验证集重新划分为新的训练和测试集而构建的，从而使相关性 每个拆分之间问题类型和答案之间的差异也不同。 例如，“网球”是在训练集中以“什么运动...”开头的问题的最常见答案，而“滑雪”是在测试集中这些问题的最常见答案。 因为训练数据中很典型而选择答案的模型在此测试集上的表现会很差。

**Bias-Only**Model:VQA-CP附带有题为65个问题类型之一的问题，与问题的前几个单词相对应（例如，“什么颜色”）。 纯偏差模型使用此分类标签作为输入，并在与主模型相同的多标签目标上进行训练。

**Main Model**我们使用BottomUpToDown（Anderson等人，2018）VQA模型的流行实现。 该模型使用多标签目标，因此我们通过将每个可能的答案视为两类分类问题来应用集成方法。

Results：表3显示了结果。 learned-mixin法非常有效，将VQA-CP的性能提高了约9个点，而熵正则化函数又可以将其提高3个点，大大超过了以前的工作。 对于learned-mixin集合，我们发现g(x_i)与偏差的预期准确性密切相关，在测试数据上的长矛相关性为0.77。 定性示例（图2）进一步表明，当模型知道是否可以依赖纯偏差模型时，模型会增加g(x_i)。

