# Generative Bias for Robust Visual Question Answering

## 简介

VQA会受到数据集分布bias的影响，导致那些数据集比例大的答案更有可能作为答案的预测结果。本文基于生成式模型，利用对抗目标和知识蒸馏对模型进行偏置调整。

## 背景

目前，非常多的研究显示VQA在数据集之间存在偏置的倾向。并且VQA非常依赖于数据集之间的语言偏置。这样会导致模型倾向于只根据问题而不是图片预测相似的回答。

针对这个问题，最近的工作利用联邦学习（集成学习）来消除这一偏置。模型会同时学习到每一个数据集或模态的偏置。比方说，在一些研究中，将QA模型作为偏置模型，利用QA通过模型回答问题给出的答案分布和问题的相关程度来决定语言先前偏置。之后，就用这个QA模型来训练一个鲁棒性高的VQA模型。如果QA模型表示先前偏置越好，那么就越可以避免VQA预测答案出现偏置的现象。

现有的基于联邦学习的方法不是用训练集进行预计算和统计，就是只计算单模态与答案之间的偏置。

本文推测目前现有的方法会出现偏置表示的限制，因为模型的表示容量会受到输入的限制。另外，预计算只能表示部分偏置，因为研究显示QA模型的标签分布和VQA模型的标签分布有显著不同。

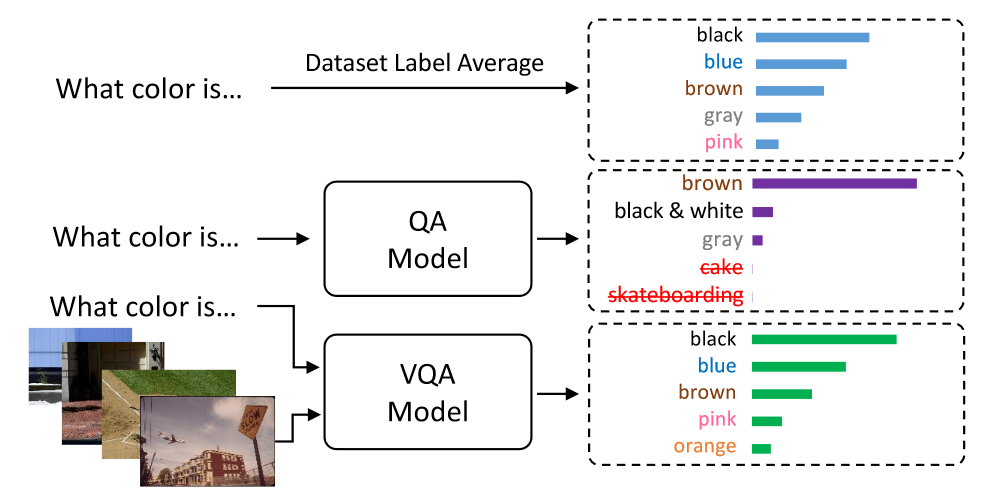


Fig1

## GenB

因此，本文提出一种新颖的、随机的偏置模型来直接从VQA模型中学习到偏置。具体而言，本文将生成式对抗模型（GAN）作为偏置模型，通过引入一个随机的噪声向量，根据给定的问题模仿VQA模型的答案分布。

由于大多数文章都采用问题中的bias，因此本文将问题作为主要的偏置模态。为了达到这一目的，本文在对抗学习之前先进行知识蒸馏，强迫偏置模型和VQA模型尽可能接近。最后，通过生成式偏置模型，本文可以修改去偏置损失函数来训练VQA模型。得到的效果比之前的方法都要好。

### VQA基线

VQA模型接收图像和问题作为一对输入，通过学习从回答集中正确预测出正确的那个回答。一个典型的VQA模型将视觉表示（*其中表示图片中物体的数量、表示向量的维度*）和问题表示作为输入。通过注意力模块和多层感知机分类器，输出一个答案分数（logit）向量，即。

这里补充一下，注意力模块的输入和输出的大小是一致的，这样能保证能在深度上延申，就像搭积木一样。

得到之后，在应用sigmoid函数，得到一组概率，并通过学习尽可能让其靠近真实的概率。

### 偏置模型联邦学习

在联邦学习方法中，存在一个偏置模型，生成输出为，还有一个目标VQA模型。在测试阶段，我们舍弃只使用。为了消除偏置，本文需要尽可能将拟合偏置。

之后，拟合好之后，目标VQA模型会用一个去偏置损失函数进行训练来提升目标VQA模型的鲁棒性。最后，目标模型通过避免给出和偏置模型一样的偏置结果，学习预测出无偏置的答案。的结构可以和一样也可以不一样，并且可以是由多个模型共同组成的。但是，由于本文认为只用单独的模态QA模型对偏置的表达能力有限，因此本文想让偏置模型和结构一样，都使用VQA模型。

### 生成式偏置

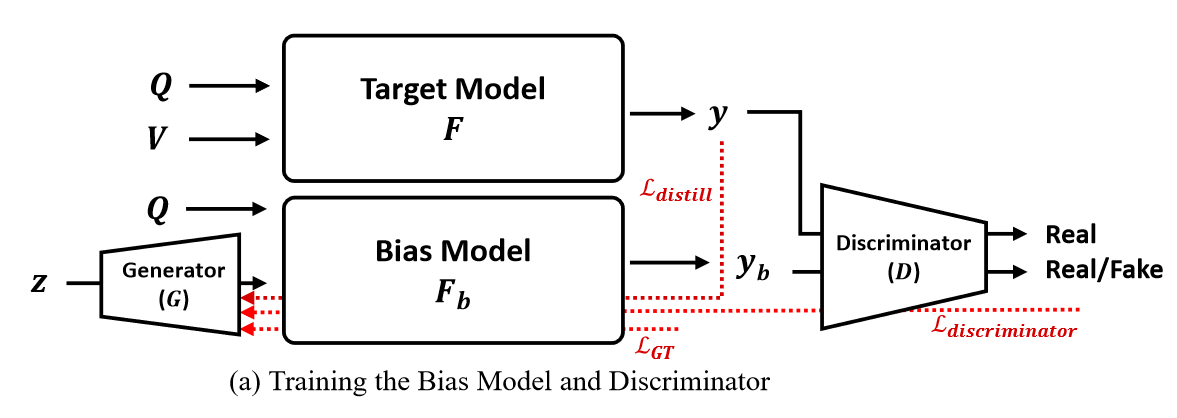


Fig2

对于偏置模型，输入的问题没有变化，但是输入的图像是一个随机噪声向量经过生成网络合成后的图片，即(噪声服从标准正态分布)。

之后，根据问题和输入的噪声图像，得到了对应分数，即。需要注意的是，也可以将输入换成原来的图像，但是通过实验表明好像没什么用。适当简化一下，将偏置模型和生成网络看成一个网络，即。

### 训练偏置模型

本文使用传统的VQA损失函数：二元交叉熵损失函数，即。

另外，为了让偏置模型捕捉到目标VQA模型的偏置，本文引入了对抗训练。具体来说，本文引入一个鉴别器，将目标VQA模型的答案预测为真，将偏置模型的答案预测为假。整体训练的损失函数如下所示：

不过损失函数并不是要越小越好，生成器（）希望损失函数越小越好，而鉴别器（）希望损失函数越大越好。这样训练之后，和的答案分布就会非常接近。

除此之外，本文还用知识蒸馏目标函数用目标VQA模型来训练偏置模型。通过实验证明KL散度最适合描述偏置模型输出和目标VQA模型输出两者之间的距离。因此增加了一个新的损失函数：

整体的损失函数即为。偏置函数想最小化这一损失函数，而鉴别器想最大化这一损失函数。

## 对目标VQA模型去偏置

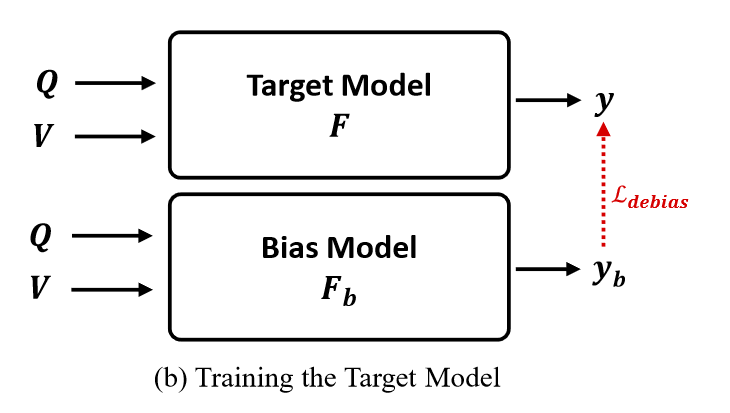


Fig3

本文设计了一个损失函数：

其中，第个元素如下进行表示：

分别代表正确答案和偏置模型的答案。

## 实验结果

对于数据集VQA-CP2、VQA-CP1来说，GenB取得了最好或次好的成果。

消融实验显示：

* 在训练偏置模型中，将三个函数都考虑进损失函数比较好。
* 本文自己定义的去偏置损失函数取得了最好的成果。
* 在选择偏置模型方面，本文的问题+噪声获得了更好的效果。
* 用GenB+其他骨干网络可以获得SOTA的准确率。

量化结果：

通过将不同的噪声输入偏置模型，可以得到不同的注意力位置。目标VQA模型能准确锁定位置，并输出正确答案。

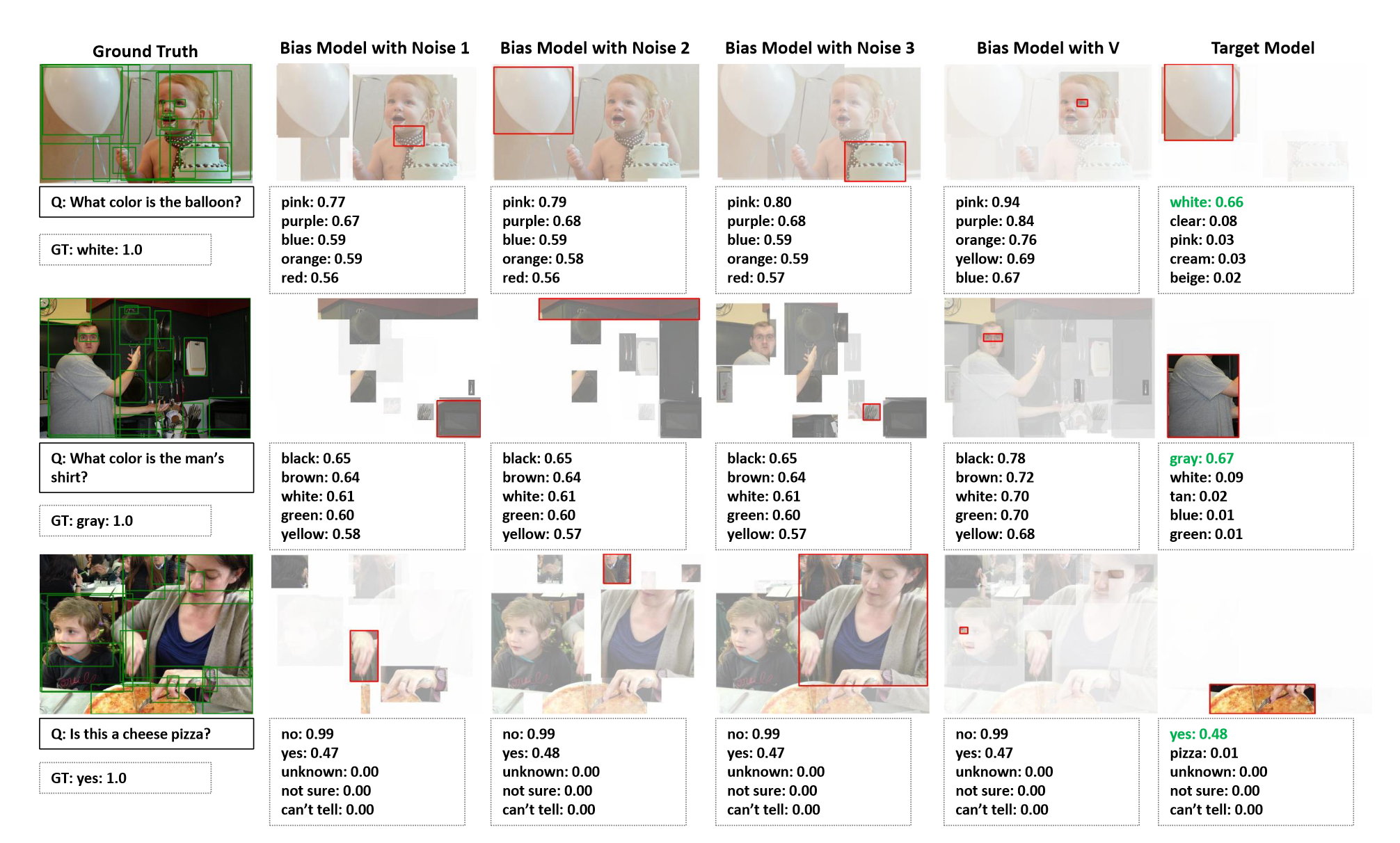


Fig4

## 个人感想

对抗学习那部分非常震撼我，整体感觉就像是瞎摸出来的，但是获得效果很好。可以学习先实验再说明的思路。

## KL 散度