Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering

Aishwarya Agrawal∗, Dhruv Batra‡, Devi Parikh‡, Aniruddha Kembhavi∗
∗Virginia Tech, ‡Georgia Institute of Technology, †Allen Institute for Artificial Intelligence
aish@vt.edu, {dbatra, parikh}@gatech.edu, anik@allenai.org

1. Introduction

Automatically answering questions about visual content is considered to be one of the holy grails of artificial intelligence. Visual Question Answering (VQA) poses a rich set of challenges spanning various domains such as computer vision, natural language processing, knowledge representation, and reasoning. In the last few years, VQA has received a lot of attention – a number of VQA datasets have been curated [1,9] and a variety of deep-learning models have been developed [10–27].

However, a number of studies have found that despite recent progress, today’s VQA models are heavily driven by superficial correlations in the training data and lack sufficient visual grounding [8,28,29]. Intuitively, it seems that when faced with a difficult learning problem, models resort to latching onto the language priors in the training datasets to the point of ignoring the image – e.g., overwhelmingly replying to ‘how many X?’ questions with ‘2’ (irrespective of X), ‘what color is …?’ with ‘white’, ‘is the . . . ?’ with ‘yes’. One reason for this emergent dissatisfactory behavior is the fundamentally problematic nature of IID train-test splits in the presence of strong priors. In existing datasets, strong priors present in the training data also carry over to the test data. As a result, models that rely on a sort of sophisticated memorization of the training data demonstrate acceptable performance on the test set. This is problematic for benchmarking progress in computer vision and AI because it becomes unclear what the source of the improvements is and to what extent VQA models have learned to ground concepts in images, and consequently significantly outperforms existing VQA models.

To help disentangle these factors, we present a new split of the VQA v1.0 dataset [1] – VQA-CP v1.0 – that enables stress testing VQA models under mismatched priors between train and test. VQA-CP v1.0 is created such that the distribution of answers per question-type (‘what color is the . . . ?’, ‘is the person’) is by design different in the train split (majority answers: ‘white’, ‘no’), compared to the test split (majority answers: ‘black’, ‘yes’). Existing VQA models tend to largely rely on strong language priors in existing datasets and suffer significant performance degradation on VQA-CP v1.0. We propose a novel model (GVQA) that explicitly grounds visual concepts in images, and consequently significantly outperforms existing VQA models.

we report the performance of several existing VQA models [11,14,17,21,30] on VQA-CP v1.0. Our key finding is that the performance of all tested existing models drops significantly when trained and evaluated on VQA-CP v1.0 compared to training and evaluation on VQA v1.0 dataset.

Our primary technical contribution is a novel Grounded Visual Question Answering model (GVQA) that contains inductive biases and restrictions in the architecture specifically designed to prevent it from ‘cheating’ by primarily relying on priors in the training data. GVQA is motivated by the following intuition – questions in VQA provide two key pieces of information:

(1) What should be recognized? Or what visual concepts in the image need to be reasoned about to answer the question (e.g., ‘What color is the plate?’ requires looking at the plate in the image).

(2) What can be said? Or what the space of plausible answers is (e.g., ‘What color . . . ?’ questions need to be answered with names of colors).
2. Visual Question Answering under Changing Priors (VQA-CP v1.0)

In VQA-CP v1.0, the distribution of answers for a given question type is by design different in train and test splits (Fig. 2), unlike VQA v1.0 dataset where the distribution for a given question type is similar across train and val splits [1]. For instance, in VQA-CP v1.0, ‘tennis’ is the most frequent answer for the question type ‘what sport’ in train split whereas ‘skiing’ is the most frequent answer for the same question type in VQA-CP v1.0 test split. Similar differences can be seen for other question types as well – ‘what animal’, ‘what color’, ‘how many’, ‘what brand’. Please visit our poster to learn more about VQA-CP v1.0.

To show the difficulty of VQA-CP v1.0, we train and evaluate the following VQA models on VQA-CP v1.0 and compare with their performance on VQA v1.0 – 1) the model [30] from the VQA paper [1] which we refer to as d-LSTM+n-I, 2) The Neural Module Networks (NMN) [14] which are designed to be compositional in nature, 3) Stacked Attention Networks (SAN) [11] which is one of the widely used models for VQA, and 4) Multimodal Compact Bilinear Pooling (MCB) [21] which won the VQA Challenge 2016. From Table 2 we can see that the performance of all the existing VQA models drops significantly in the VQA-CP v1.0 setting compared to the VQA v1.0 setting.

Table 2. Accuracies of existing VQA models on the VQA v1.0 val split when trained on VQA v1.0 train split and those on VQA-CP v1.0 test split when trained on VQA-CP v1.0 train split.

3. Grounded Visual Question Answering (GVQA) Results

Table 1 shows the performance of the GVQA model in comparison to several existing VQA models on the VQA-CP v1.0 dataset using the VQA evaluation metric [1]. It is insightful to look at the accuracies of ‘Yes’ and ‘No’ individually because the distribution of ‘Yes’ and ‘No’ in the test set is biased towards ‘Yes’ (74.64%). Hence, a model which says ‘Yes’ for all Yes/No questions, would score 100% for Yes but perform poorly on No and normalized Yes/No (Norm Y/N) accuracies.

As shown in Table 1 our proposed GVQA significantly outperforms past VQA models on the VQA-CP v1.0 dataset. In particular, GVQA does much better on Y/N questions. Interestingly, GVQA also outperforms MCB on the overall metric as well as the Yes/No category, in spite of the MCB model using the more powerful ResNet-152 architecture in its image pipeline, compared to VGG-16 in GVQA. Please visit our poster to learn about GVQA.

Table 1. Accuracies of our model compared to existing VQA models on the VQA-CP v1.0 dataset.

Table 2. Accuracies of existing VQA models on the VQA v1.0 val split when trained on VQA v1.0 train split and those on VQA-CP v1.0 test split when trained on VQA-CP v1.0 train split.
References


