An Empirical Evaluation of Visual Question Answering for Novel Objects

Santhosh K. Ramakrishnan[1,2] Ambar Pal[1] Gaurav Sharma[1] Anurag Mittal[2]
[1]IIT Kanpur[*] [2]IIT Madras[†]

Abstract

We study the problem of answering questions about images in the harder setting, where the test questions and corresponding images contain novel objects, which were not queried about in the training data. Such setting is inevitable in real world—owing to the heavy tailed distribution of the visual categories, there would be some objects which would not be annotated in the train set. We show that the performance of two popular existing methods drop significantly (up to 28%) when evaluated on novel objects cf. known objects. We propose methods which use large existing external corpora of (i) unlabeled text, i.e. books, and (ii) images tagged with classes, to achieve novel object based visual question answering. We do systematic empirical studies, for both an oracle case where the novel objects are known textually, as well as a fully automatic case without any explicit knowledge of the novel objects, but with the minimal assumption that the novel objects are semantically related to the existing objects in training. The proposed methods for novel object based visual question answering are modular and can potentially be used with many visual question answering architectures. We show consistent improvements with the two popular architectures and give qualitative analysis of the cases where the model does well and of those where it fails to bring improvements.

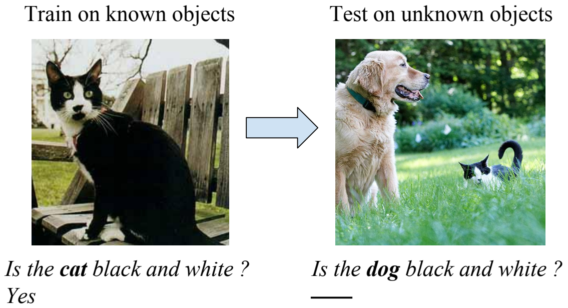
Train on known objects
Test on unknown objects
Is the **cat** black and white ?
Yes
Is the **dog** black and white ?

Figure 1: We are interested in answering questions about images containing objects not seen at training.

1. Introduction

Humans seamlessly combine multiple modalities of stimulus, e.g. audio, vision, language, touch, smell, to make decisions. Hence, as a next step for artificial intelligence, tasks involving such multiple modalities, in particular language and vision, have attracted substantial attention recently. Visual question answering (VQA), i.e. the task of answering a question about an image, has been recently introduced in a supervised learning setting [21, 3]. In the currently studied setup, like in other supervised learning settings, the objects in the training data and the test data overlap almost completely, i.e. all the objects that appear during testing have been seen annotated in the training. This setting is limited as this requires having training data for all possible objects in the world—this is an impractical requirement owing to the heavy tailed distribution of the visual categories. There are many objects, on the tail of the distribution, which are rare and annotations for them might not be available. While humans are easily able to generalize to novel objects, e.g. make predictions and answer questions about a wolf, when only a cat and/or a dog were seen during training, automatic methods struggle to do so. In the general supervised classification, such a setting has been studied as *zero shot learning* [15], and has been applied for image recognition as well [11, 14, 36, 40]. While the zero shot setup works with the constraint that the test classes or objects were never seen during training, it also assumes some form of auxiliary information to connect the novel test classes with the seen train classes. Such information could be in the form of manually specified attributes [11, 14, 40] or in the form of relations captured between the classes with learnt distributed embeddings like, `Word2Vec` [23] or `GloVe` [25], of the words from an unannotated text corpus [36]. In the present paper, we are interested in a similar setting, but for the more unconstrained and challenging task of answering questions about novel objects present in an image. Such a setting, while being natural, has not been studied so far, to the best of our knowledge.

We start studying the problem by first proposing a novel split (§4.1), into train and test sets, of the large-scale pub-

[*] The project started when Santhosh Ramakrishnan and Ambar Pal were summer interns at IIT Kanpur. Ambar Pal is a student at IIIT Delhi. `ambar14012@iiitd.ac.in, grv@cse.iitk.ac.in`

[†] `{ee12b101@ee, amittal@cse}.iitm.ac.in`