Inductive Biases for Low Data VQA: A Data Augmentation Approach

Narjes Askarian[1], Ehsan Abbasnejad[2], Ingrid Zukerman[1], Wray Buntine[1], Gholamreza Haffari[1]
[1]Dept. of Data Science and AI, Monash University [2]Australian Institute for Machine Learning, The Univ. of Adelaide

1. Motivations

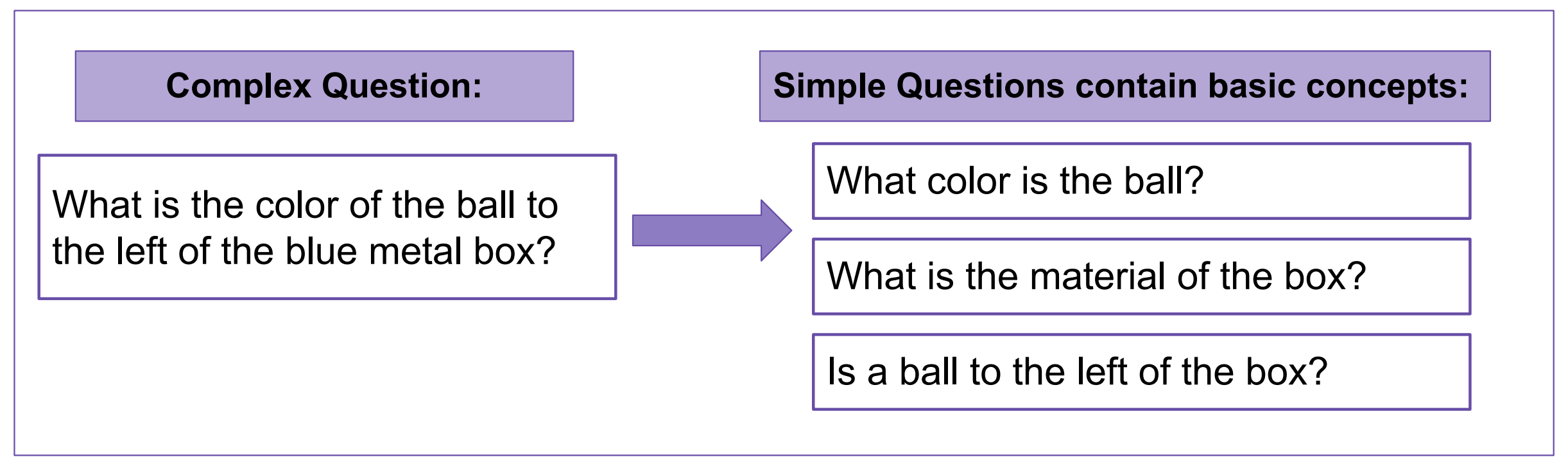
Existing VQA models have recently achieved remarkable results when training on large-scale labeled datasets. However, annotating large amounts of data is not feasible in many domains. VQA models do not maintain their performance in low data scenarios.

we address the problem of VQA in a low-labeled data regime using a data augmentation approach to enlarge the initial small labeled dataset in order to inject proper inductive biases to the QA model.

Inductive biases

An inductive bias that a typical learner acquires by training on natural language tasks is related to the inherent compositionality of the human language, *e.g.*, a complex sentence can be understood by understanding its simpler chunks.

VQA datasets mostly contain complex questions. Complex questions can be learned on the basis of the basic concepts. We hypothesized that augmenting the training set of complex questions with simpler questions will help the model to better learn the basic concepts.


Complex Question:

What is the color of the ball to the left of the blue metal box?

Simple Questions contain basic concepts:

What color is the ball?

What is the material of the box?

Is a ball to the left of the box?

2. Data Augmentation Method

Question simplicity

Based on the focus of this work, which is reasoning using compositionality, we represent simplicity as the number of reasoning steps required to answer a question.

Question Ambiguity

A question is considered ambiguous when the target object cannot be uniquely identified by the attributes stated in the question.

Creating Unique attribute combinations

Unique attribute combinations, uacs, are referring expressions that consist of a set of attributes which can uniquely identify an object in an image. We use image scene information include in many VQA datasets to create uacs. We start with creating combinations of length 1 for all objects in the image, $uac_{l=1}$ by simply comparing the values of the same attributes in all objects of the image, *e.g.,* the colors or shapes of the objects; and then record the attribute values that appear in only one object such as"red"in the figure.

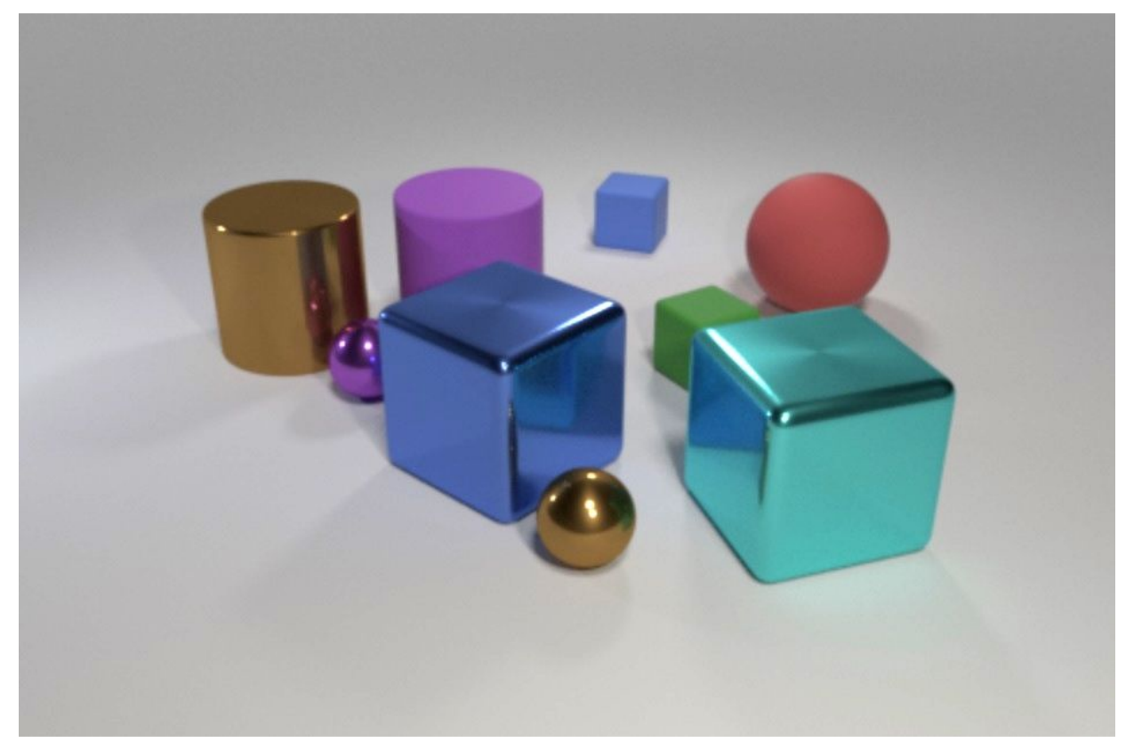


Figure 1: The question *"what color is the object behind the blue cube?"* is ambiguous since there are two blue cube in the image. To create an unambiguous question, the target object must be uniquely identifiable by the referring expression in the question. A unique attribute combination assures that the expression refers to a unique object such as ***"red"*** or ***"red sphere"*** that both refer to the red sphere at the top right corner.

Generating template-based questions

We synthesize the templates for two types of questions: *query-attribute* and *existential* as can be seen in the following table. [attribute] indicates a placeholder which must be replaced with the queried attribute name that can be any of the attributes from the attribute set, *i.e.*, *size, color, material, shape*, while [uac] indicates a uac of the desired object.

Question Type	Template	Example
Query-attribute	What [attribute] is the [uac] object?	What size is the red object?
Existential	Is there a [uac] object?	Is there a red object?

3. Results

To simulate a low-data scenario, we select four small subsets from the CLEVR training set [1] with different sizes denoted as a percentage of the full dataset referred to as We *s-CLEVR{x}* where *x∊ {5,10,20, 30}* indicates the size of the subset. We use the scene information of the images from *s-CLEVR{x}* for generating augmented questions. We evaluate our approach using the modular VQA model proposed in [2].

The experiments are conducted in two stages: ***data augmentation*** and ***training***.

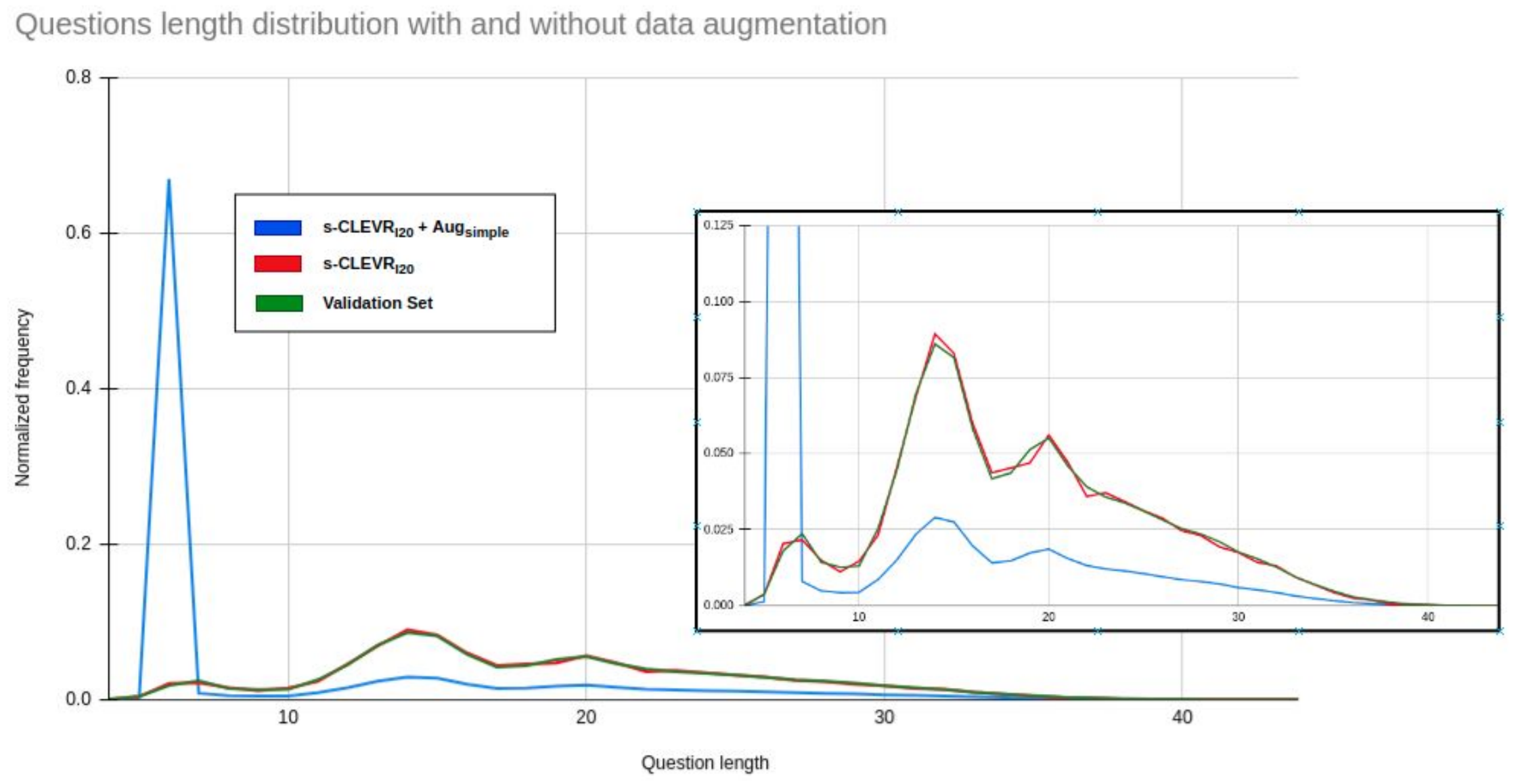
- The first stage includes extracting uacs and generating question-answer pairs using the proposed data augmentation method called Aug^{simple}.
- In the second stage, we add the augmented questions to the training set to create an augmented training set, *i.e.*, *s-CLEVR{x}*+Aug^{simple}.

Data augmentation statistics

This table shows the statistics of generated questions for each training subset. Specifically, it shows the length and number of generated questions per question type and in total.

	Exist Questions	Attribute Questions	All Questions
Subcategories	yes, no	color, shape, size, material	-
Questions length	6	6	6
Program Length	4	4	4
Count in 5%	10,817 (16%)	57,996 (84%)	68,813
Count in 10%	21,715 (16%)	115,568 (84%)	137,283
Count in 20%	43,375 (16%)	230,824 (84%)	274,199
Count in 30%	65399 (16%)	347,304 (84%)	412,703

This diagram shows a comparison of the distributions of questions length with and without data augmentation. *s-CLEVR{20}* and val set are sampled from the dataset original distribution while *s-CLEVR{20}*+Aug^{simple} shows the distribution of the augmented training set.



Training results

Training Set	5%	10%	20%	30%
Aug^{simple}	31.69	30.18	30.14	30.17
s-CLEVR	46.91	49.90	54.24	87.70
s-CLEVR$_+$Aug^{simple}	69.23	83.06	87.81	91.26

References:

[1] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. *Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2901–2910.

[2] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. *Inferring and executing programs for visual reasoning.* In IEEE International Conference on Computer Vision (ICCV), pages 2989–2998.