# Inductive Biases for Low Data VQA: A Data Augmentation Approach

**Narjes Askarian**
Dept. of Data Science and AI
Monash University

**Ehsan Abbasnejad**
Australian Institute for Machine Learning
The Univ. of Adelaide

**Ingrid Zukerman** and **Wray Buntine** and **Gholamreza Haffari**
Dept. of Data Science and AI
Monash University

VQA is the task of answering questions about visual contexts, which has attracted significant attention and achieved impressive results (Hudson and Manning, 2019b; Teney et al., 2017; Agrawal et al., 2018). This success is partly due to the use of large-scale labeled datasets (Hudson and Manning, 2019a; Antol et al., 2015; Zhu et al., 2016; Johnson et al., 2017). Relying on large-scale labeled datasets is not realistic in many settings, due to the infeasibility of collecting such data. In addition, the objective of VQA is rather ambitious, as there is potentially an infinite number of questions to be asked about an image scene. We therefore argue that VQA is a problem caused by low data, i.e., in the absence of sufficient data, current VQA systems do not maintain their high performance (e.g., see Figure 1).

We consider VQA in low-labeled data scenarios, and investigate the features of a VQA task that necessitate a lot of labeled training data. One of these features is understanding complex questions about rich visual contexts.VQA datasets mostly contain complex questions where a learner must identify multiple objects and understand their relationships. Understanding a question and capturing an image scene is a lot easier when the learner has access to a large amount of labeled data. The model eventually captures the complexity of a situation after seeing a wide variety of data when training on a large dataset. However, complex relationships are challenging to learn from small datasets.

In this paper, we improve the generalisation of VQA models by injecting inductive biases, so that the model can explicitly have access to them in a data-efficient manner. These inductive biases heavily impact the answers in VQA. 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. Normally, it is easier to capture the meaning of the resulting chunks than the original sentence, which provides a powerful foundation for understanding complex sentences.

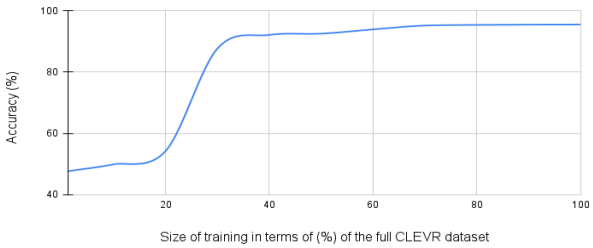


Figure 1: Accuracy of vanilla training of the *execution engine* on CLEVR `val` when trained on different-sized random subsets of the CLEVR `train` set.

Inspired by the fact that a complex question can be learned from its basic concepts, we hypothesized that augmenting the training set of complex questions with simpler questions will help the model. The notion of question simplicity can be defined on the basis of different criteria, including syntactic and semantic dimensions. In the VQA context, we consider simplicity as the number of reasoning steps required to answer a question. Thus, the simplest possible question requires identifying a single object and reasoning about it.In particular, we include simple questions that, if learned, could lead to better representations in the VQA model.

We take a data augmentation approach and enlarge the initial small training set by automatically generating simple question-answer pairs for images. The idea is that basic concepts can be learned from simple questions, enabling the model to better learn the structure of more complex questions.

Data augmentation strategies have proven to be particularly useful in a variety of computer vision applications, including image classification (Krizhevsky et al., 2012). Not only can they be helpful in overcoming the problem of insufficient

labeled data, they are also used to reduce overfitting and class imbalance problems (Shorten and Khoshgoftaar, 2019). Current data augmentation techniques use data warping or oversampling to increase the size of the training dataset (Ruprecht and Muller, 1995; Shorten and Khoshgoftaar, 2019). Data warping is a technique for transforming data while maintaining its labels. Typically the examples are transformed by geometric and color transformations, random erasing, neural style transfer, and adversarial training.

Data augmentation in VQA is under-explored due to the challenge of correctly preserving the semantic relation of the *<image, questions, answer>* triplet during transformation. Geometric transforms and random cropping of the image cannot guarantee the preservation of the answer. For instance, the answer to *"What color is the thing on the left side of the cube?"* may be flipped if the image is vertically transformed. Random cropping can result in missing the number of objects when counting to answer a *how many* question.

Our proposed data augmentation method automatically generates simple questions. Our method only requires having access to shallow annotations of an image, and does not use any additional labeled data. These annotations give some information about the appearance of the objects in the image. The answers to the questions are also automatically generated at no cost in human effort. The method is generic and is applicable to any VQA task given the scene information available in many current VQA datasets. The statistics shows that, with respect to the distribution of the questions lengths in the original dataset, our augmented training set dramatically change the distribution in favour of shorter questions. The experimental results and analysis demonstrate that our method is effective in improving VQA performance, yielding an improvement in accuracy of up to 34% compared to training on only the initial labeled data. Table

## References


Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*, pages 4971–4980.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Drew A Hudson and Christopher D Manning. 2019a. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Drew A. Hudson and Christopher D. Manning. 2019b. Learning by Abstraction: The Neural State Machine. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5901–5914.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, Red Hook, NY, USA. Curran Associates Inc.

D. Ruprecht and H. Muller. 1995. Image warping with scattered data interpolation. *IEEE Computer Graphics and Applications*, 15(2):37–43. Conference Name: IEEE Computer Graphics and Applications.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60.

Damien Teney, Lingqiao Liu, and Anton Van Den Hengel. 2017. Graph-Structured Representations for Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3233–3241, Honolulu, HI. IEEE.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004.