Supplementary Material for: Visual Concept-Metaconcept Learning

Chi Han*
MIT CSAIL and IIIS, Tsinghua University

Jiayuan Mao*
MIT CSAIL

Chuang Gan
MIT-IBM Watson AI Lab

Joshua B. Tenenbaum
MIT BCS, CBMM, CSAIL

Jiajun Wu
MIT CSAIL

1 Dataset Visualization

In this section, we explain in detail the training-test split for the datasets with examples, with images from CLEVR[Johnson et al., 2017], GQA[Hudson and Manning, 2019] and CUB[Wah et al., 2011].

In this paper we consider two perspectives of generalization. First, we demonstrate that, with the help of metaconcept questions, VCML is capable of improving its data efficiency in concept learning as well as generalizing to unseen visual attributes compositions.

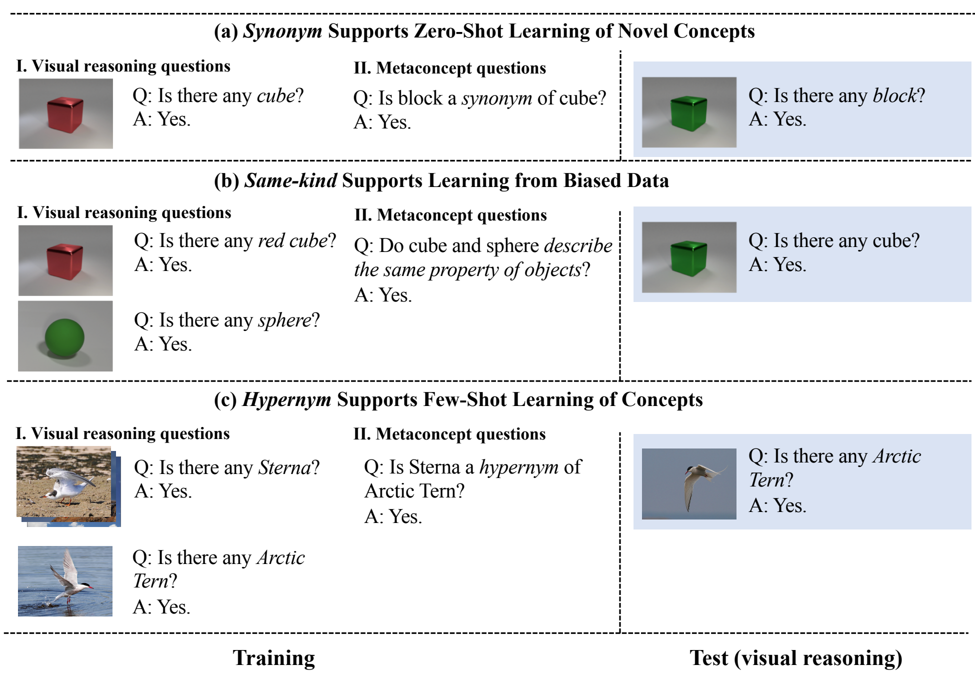

Figure 1: Data split used for tests on "Metaconcepts help concept learning". In training, the visual data are insufficient or even biased for some concepts. The models are then required to leverage the information provided in metaconcept questions in order to correctly reason on visual questions about these concepts (shown in blue).

First two authors contributed equally. Work was done when Chi Han was a visiting student at MIT CSAIL. Project Page: `http://vcml.csail.mit.edu`.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Figure 2: Data split used for metaconcept generalization tests. The models are required to leverage the visual analogy between concepts to predict metaconcepts about unseen pairs of concepts (shown in blue). See the main text for details.

Figure 1 shows examples for the three tasks evaluated in this type of generalization. In training, we use visual reasoning datasets for learning the visual grounding of concepts. Furthermore, we provide metaconcept questions to provide abstract-level supervision. In test (shown in blue rectangles), the models are required to correctly reason on visual concepts.

Second, we study how learned visual concepts provide grounding cues to predict metaconcept relations between unseen pairs of concepts. Figure 2 illustrates the training-test splits for four metaconcept generalization tests. In training, we provide visual reasoning questions and a subset of metaconcept questions. In test, the models are required to generalize the learned metaconcepts to unseen pairs of concepts.

2 Ablation: *same_kind* Supports Learning from Biased Data

To further quantify the effectiveness of metaconcepts in supporting learning from biased data, we conducted an ablation study. Recall that the training questions are from two sources: all images in the split A, and a small number of images from the split B. We plot the performance of different models by varying the number of training images from the split B.

Three models are tested in Figure 3: VCML, NS-CL [Mao et al., 2019], MAC [Hudson and Manning, 2018]. We also evaluate the performance of VCML if all metaconcept questions are absent, shown as VCML (ablation).

Overall, VCML outperforms two other baselines by a margin when the number of training images from the split B is greater than 3. NS-CL significantly closes the gap when there are more than 10 split B images. However, VCML trained without metaconcept questions also achieves a comparable performance when the number of split B images is large. This suggests that our model outperforms both baselines in utilizing metaconcepts to support learning from biased data.

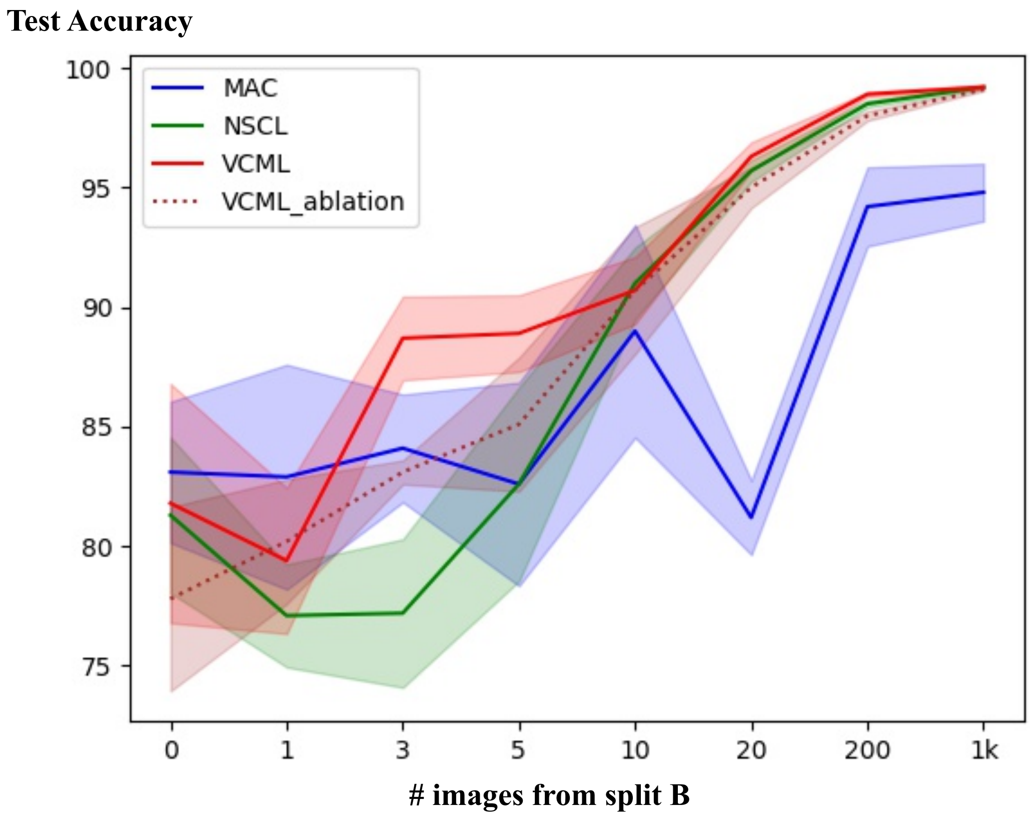

Figure 3: Models results under different levels of visual compositional bias in the experiment "*same_kind* Supports Learning from Biased Data". The y-axis is the test accuracy in percent, and the x-axis is the number of split B images in training. Plots are results for different models/settings. The transparent band denotes $\pm\frac{1}{2}$ standard deviation.

References

Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018. 2

Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Compositional Question Answering over Real-World Images. *CVPR*, 2019. 1

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *ICLR*, 2019. 2

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1