Characterizing how Visual Question Answering models scale with the world

Eli Bingham*
Uber AI Labs
San Francisco, CA 94103
eli.bingham@uber.com

Piero Molino*
Uber AI Labs
San Francisco, CA 94103
piero@uber.com

Paul Szerlip*
Uber AI Labs
San Francisco, CA 94103
pas@uber.com

Fritz Obermeyer
Uber AI Labs
San Francisco, CA 94103
fritzo@uber.com

Noah D. Goodman†
Department of Psychology
Stanford University
Stanford, CA 94305
ngoodman@stanford.edu

Abstract

Detecting differences in generalization ability between models for visual question answering tasks has proven to be surprisingly difficult. We propose a new statistic, *asymptotic sample complexity*, for model comparison, and construct a synthetic data distribution to compare a strong baseline CNN-LSTM model to a structured neural network with powerful inductive biases. Our metric identifies a clear improvement in the structured model's generalization ability relative to the baseline despite their similarity under existing metrics.

1 Introduction

People are exposed to a wide variety of visual situations, but this variety is still impossibly small relative to all the possible combinations of events and objects that could occur in the real world. Humans are nevertheless capable of recognizing and behaving in unusual situations without any undue difficulty.

Good visual question answering models ought to be similarly flexible. However, determining whether one model represents genuine progress over another has not been straightforward [6, 8].

One reason for this is dataset bias [26]: simple models that can exploit dataset statistics as a shallow source of commonsense knowledge can exhibit surprisingly high performance when evaluated in standard ways [3]. VQA research is especially vulnerable to dataset bias due to (necessarily) complicated dataset creation processes with open-ended annotator prompts. In natural images selected by humans as interesting or salient (e.g. Flickr images) some types of objects appear much more frequently than others. Moreover, most types of objects almost never appear together; even when they do, human annotators asked to quickly list a few salient relationships per image may unconsciously favor certain kinds of relationships over others [15].

To measure model behavior in unusual situations and expose dataset adaptation, the metric used to evaluate a model (e.g. top-1 accuracy) can be computed on a test dataset reweighted by the inverse of

*Equal contribution. These authors listed in alphabetical order.
†Uber AI Labs, ndg@uber.com

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

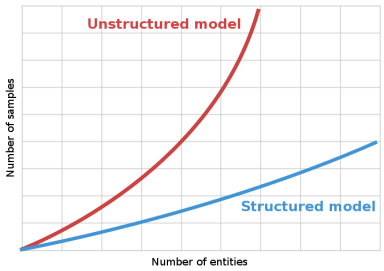

Figure 1: Hypothesis: models that incorporate syntactic and semantic prior knowledge exhibit much less dramatic asymptotic sample complexity

some measure of data frequency (e.g. the product of normalized empirical object frequencies):

$$x, y \sim p_{\mathcal{D}}^{reweighted}(x, y) = \frac{1}{\#(x, y)} p_{\mathcal{D}}(x, y) \tag{1}$$

Unfortunately, evaluation with reweighted data may not be enough in practice to distinguish between models. That is, the new metric does not necessarily capture the relationship between the complexity of the data distribution (e.g. the number of object types) and the dataset size, which is often determined by exogenous factors like cost and collection time. Two models with equal reweighted accuracy on a given dataset may perform differently on a dataset with twice the number of objects or half the number of samples.

This argument suggests that the true measure of interest is the rate at which the number of samples required for a model to achieve a given accuracy level on reweighted data increases as the complexity of the data distribution grows, the *asymptotic sample complexity*.

The hypothesis is that better models for VQA should "scale with the world (as in Figure 1)." For models with better asymptotic sample complexity, as the underlying complexity of the data increases, less additional data points are required to maintain the same level of performance.

To test this hypothesis, asymptotic sample complexity is estimated by training two VQA models on a synthetic data distribution that reflects statistics extracted from the Visual Genome (Section 6.2). The two models, a strong baseline recurrent neural network model that is competitive with state-of-the-art models on real world VQA datasets [8] and a simple structured neural network model, serve as a stand-in for the unstructured and structured models of the hypothesis, respectively.

2 Data

2.1 SimTown

In recent years, there has been a proliferation of publicly available synthetic datasets [28, 27, 9, 7, 31], which are important tools in grounded language understanding research. Directly testing the hypothesis that two models with the same metric performance can exhibit different asymptotic behavior requires a generative process with easily adjustable mechanisms to modify both complexity and dataset size.

To implement these complexity levers, this paper introduces a custom simulated environment built in Unity3D called SimTown (Section 6.4). SimTown is a cartoon 3D town in which roads, sidewalks, trees, buildings, cars, pedestrians and other objects can be placed, configured, and moved programmatically. Additionally, the environment can extract high-level semantic and visual ground-truth information from scenes as specified within a grounded language understanding task.

2.2 Generating synthetic visual question answering data

In effect, SimTown enables the construction of a simple generative model of visual question answering datasets with clear control over the richness of the visual and linguistic world.

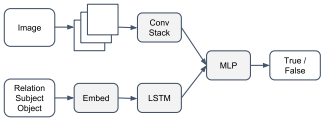

Figure 2: The unstructured relationship verification model

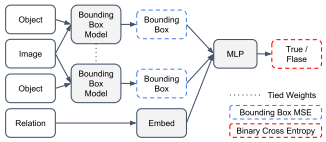

Figure 3: A high-level overview of PBN, a structured spatial relationship verification model

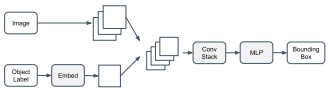

Figure 4: The PBN bounding box module

To isolate the problem of generalizing to low-frequency questions and scenes, two strong constraints are placed on the linguistic structure.

First, synthetic questions are restricted to a simple form: yes/no questions that represent a single binary relationship, like "is a red truck left of a green car (in this scene)?" While simple, understanding binary relationships is a key challenge in grounded language understanding [16, 4, 30, 32, 20]

Second, an analysis of the Visual Genome (VG) dataset (Table 2) shows that spatial relationships are both the simplest and by far the most common type of relationship present in the VG relationship corpus. Thus, a restriction is made to planar spatial relationships, e.g. "left of" and "right of."

Further restrictions are placed on the types of images generated in SimTown, with street scenes always containing either two or three different vehicles placed in random, non-overlapping locations on a two-way road. The third vehicle is intended to distract the model and make the task more difficult. To simulate additional visual variation, a random row of buildings with mailboxes and trees are included in the background, along with random weather conditions.

Within a dataset, each example is a tuple of an image, question, and answer. To generate a single valid tuple, rejection sampling cycles through potential image, question, answer pairs until a scene matches the linguistic description.

In detail, the process first samples a 3D scene and an answer (true or false), renders an image from the scene at a fixed location camera, then generates questions until the question's logical form evaluated on the latent scene information (ground-truth vehicle positions and types) matches the answer. All datasets have an equal number of true and false examples.

Rather than choosing the types of the vehicles in training and validation datasets uniformly at random, the frequency distribution of the k most common entities from the VG relationship corpus is computed, and that is set as the frequency distribution for the vehicle types (see Section 6.3 for details). Note, as an approximation to evaluating the model on reweighted accuracy (equation 1), the test set object frequency distributions are instead uniform. As a result, relationships from the tail of the distributions are likely to appear in the test sets.

3 Models

The data requirements for two different types of models are examined. The first type is an unstructured deep neural network, denoted as CNN-LSTM, whose architecture is a simplification of several strong baseline models for visual question answering [18, 32]. As [6, 8] show, with a carefully specified objective, these models achieve performance parity with the more complex models described in e.g. [14, 2, 29], so CNN-LSTM performance in SimTown should proxy this line of work.

The common feature of the class of models represented by the CNN-LSTM is their complete lack of prior knowledge outside of the use of convolutional sub-networks. However, there are two structural properties of spatial relationships in SimTown that could be incorporated to make a model more effective. Syntactically, relations are made up of two objects which may be drawn from same universal set and a predicate which is a function of the two objects. Furthermore, the semantics of spatial relationships depends explicitly on the spatial properties, such as pose and position, of their constituent objects.

The second type of models is a simple structured neural network model, which we call a Place-Binder Network (PBN), that is related in spirit to the *relationship modules* in [13] and [20] and relation

networks [22]. The PBN architecture 3 is a straightforward exploitation of these structural properties of spatial relationships. To exploit the syntactic property, PBN shares object information across both relationship slots through weight-tying and computes the final truth function directly on estimates of the two objects' positions (inspired by recent work on deep models with explicit object variables [1, 17] and on CNNs for object detection [21]). Along with the objects' position, an embedding of the predicate label feed into the last MLP component, allowing the model to potentially exploit the spatial semantics, similar to work on relational models [19].

Within the PBN bounding box module (Figure 4), object label embeddings are concatenated to input images and processed through a conventional stack of convolutional layers followed directly by fully connected layers. Notably, by adding object label embeddings to the bounding box module, the relational query can influence the detection process.

4 Results

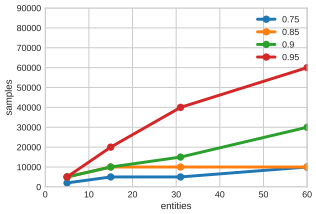

Figure 5: CNN-LSTM - Samples to achieve target performance, no distractor

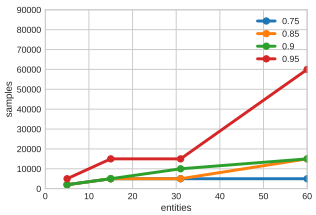

Figure 6: PBN - Samples to achieve target performance, no distractor

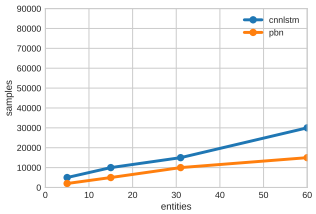

Figure 7: Samples to achieve 0.9 accuracy, no distractor

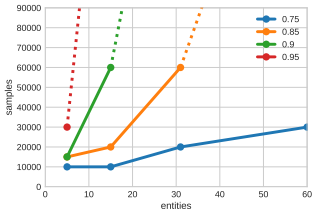

Figure 8: CNN-LSTM - Samples to achieve target performance, with distractor

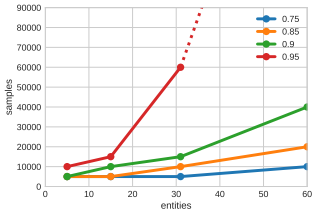

Figure 9: PBN - Samples to achieve target performance, with distractor

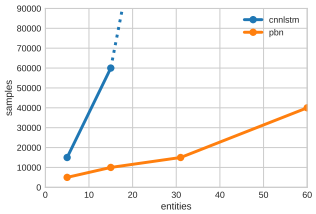

Figure 10: Samples to achieve 0.9 accuracy, with distractor

The experiments here are intended to directly estimate the curves in our hypothesis (Figure 1): how much data is necessary for the CNN-LSTM and PBN models to achieve a given level of reweighted accuracy? To estimate this measure, SimTown datasets are generated with each of 5, 15, 31, and 60 entities and 2,000, 5,000, 10,000, 20,000, 30,000, 40,000, and 60,000 training samples with and without an additional distractor car on the road. Both models are trained and evaluated on each dataset, and for each number of entities, the minimum number of samples required to reach 75%, 85%, 90%, and 95% test accuracy are collected (experimental details are available in Section 6.5). Results are shown in figures 5 - 10, where dotted lines denote a model failed to achieve a target accuracy on the largest dataset (60,000 samples).

The experiments reveal multiple datasets where both models achieve at least 90% reweighted accuracy (Figure 7 and 10), yet differ dramatically in generalization ability. A model with poor asymptotic sample complexity, like the CNN-LSTM, trained only on such a dataset would look competitive on accuracy, but require much larger datasets to maintain performance for any increase in complexity!

Furthermore, comparing the asymptotic sample complexity curves for the two models clearly favors the PBN (Figure 6 and 9) over the CNN-LSTM (Figure 5 and 8). These results match our hypothesis and intuitions, where structured models can achieve the same level of accuracy with less additional

data as entity count increases. In fact, on the more difficult distractor datasets, the CNN-LSTM requires many more samples to reach the same accuracy threshold as the PBN, sometimes even failing to reach the threshold altogether.

5 Conclusion

This paper introduces *asymptotic sample complexity* to compare models for visual question answering. Despite the difficulty of identifying clear improvements in VQA generalization with conventional metrics like accuracy [6], we hypothesized asymptotic sample complexity would distinguish performance between a structured model (PBN) and an unstructured model (CNN-LSTM). To test the hypothesis, synthetic datasets are generated inside SimTown, a tool we introduce for generating synthetic grounded language understanding tasks. Models trained on SimTown datasets reveal similar accuracy scores, yet diverging asymptotic sample complexity, confirming our hypothesis.

References

[1] S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. *arXiv:1603.08575 [cs]*, March 2016. 00025 arXiv: 1603.08575.

[2] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *arXiv:1606.01847 [cs]*, June 2016. 00014 arXiv: 1606.01847.

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *arXiv:1612.00837 [cs]*, December 2016. arXiv: 1612.00837.

[4] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4418–4427. IEEE, 2017.

[5] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, February 2015. 00688 arXiv: 1502.03167.

[6] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting Visual Question Answering Baselines. *arXiv:1606.08390 [cs]*, June 2016. 00003 arXiv: 1606.08390.

[7] 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. *arXiv:1612.06890 [cs]*, December 2016. 00000 arXiv: 1612.06890.

[8] Vahid Kazemi and Ali Elqursh. Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering. *arXiv:1704.03162 [cs]*, April 2017. arXiv: 1704.03162.

[9] Douwe Kiela, Luana Bulat, Anita L. Vero, and Stephen Clark. Virtual Embodiment: A Scalable Long-Term Strategy for Artificial Intelligence Research. *arXiv:1610.07432 [cs]*, October 2016. 00000 arXiv: 1610.07432.

[10] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014. 00882 arXiv: 1412.6980.

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv:1602.07332 [cs]*, February 2016. 00069 arXiv: 1602.07332.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 00296.

[13] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual Relationship Detection with Language Priors. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, number 9905 in Lecture Notes in Computer Science, pages 852–869. Springer International Publishing, October 2016. 00001 DOI: 10.1007/978-3-319-46448-0_51.

[14] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical Question-Image Co-Attention for Visual Question Answering. *arXiv:1606.00061 [cs]*, May 2016. 00012 arXiv: 1606.00061.

[15] Willem M. Mak, Wietske Vonk, and Herbert Schriefers. The Influence of Animacy on Relative Clause Processing. *Journal of Memory and Language*, 47(1):50–68, July 2002. 00274.

[16] Mateusz Malinowski and Mario Fritz. A Pooling Approach to Modelling Spatial Relations for Image Retrieval and Annotation. *arXiv:1411.5190 [cs]*, November 2014. arXiv: 1411.5190.

[17] Mateusz Malinowski, Ashkan Mokarian, and Mario Fritz. Mean Box Pooling: A Rich Image Representation and Output Embedding for the Visual Madlibs Task. pages 111.1–111.12. British Machine Vision Association, 2016.

[18] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. pages 1–9, 2015. 00072.

[19] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, 104(1):11–33, January 2016. 00047.

[20] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *ICCV 2017-International Conference on Computer Vision 2017*, 2017.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 00397.

[22] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv:1706.01427 [cs]*, June 2017. arXiv: 1706.01427.

[23] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014. 02287 arXiv: 1409.1556.

[24] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. *arXiv:1412.6806 [cs]*, December 2014. 00023 arXiv: 1412.6806.

[25] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 01412.

[26] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, June 2011. 00466.

[27] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, and Devi Parikh. Learning Common Sense Through Visual Abstraction. pages 2542–2550, 2015. 00009.

[28] Jason Weston, Antoine Bordes, Sumit Chopra, Tomas Mikolov, Alexander M. Rush, and Bart van Merriënboer. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv:1502.05698 [cs, stat]*, February 2015. 00018 arXiv: 1502.05698.

[29] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic Memory Networks for Visual and Textual Question Answering. *arXiv:1603.01417 [cs]*, March 2016. 00029 arXiv: 1603.01417.

[30] Xuchen Yao, Jonathan Berant, and Benjamin Van Durme. Freebase QA: Information Extraction or Semantic Parsing? *ACL 2014*, page 82, 2014. 00012.

[31] Stephanie Zhou, Alane Suhr, and Yoav Artzi. Visual Reasoning with Natural Language. *arXiv:1710.00453 [cs]*, October 2017. arXiv: 1710.00453.

[32] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded Question Answering in Images. *arXiv:1511.03416 [cs]*, November 2015. 00027 arXiv: 1511.03416.

Data type	#Instances
Images	108,777
Region Descriptions	5.4M
Question-Answer pairs	1.7M
Objects	3.8M
Attributes	2.8M
Relationships	2.3M

Table 1: An overview of the Visual Genome

Spatial	#Rel	Physical	#Rel	Action	#Rel
On	707901	Sit on	15687	Have	278647
In	241004	Stand	8250	Wear	51996
Above	57000	Hang on	6105	Hold	43151
Behind	47398	Lay on	6100	Carry	5825
Next to	46000	Cover to	5316	Eat	5218
Next	27490	Park on	2728	Walk	4723
Under	19134	On back of	1926	Play	4098
Front	18000	Grow on	1097	Watch	3987

Table 2: Counts of Visual Genome relationships

6 Appendix

6.1 Visual Genome preprocessing

Visual Genome is a large scale Visual Question Answering dataset. It contains more than 100,000 images tagged with regional descriptions, objects with attributes and mutual relationships, and question-answer pairs regarding the images. Statistics of the dataset are summarized in Table 1.

The free-form nature of the data generation prompt meant that the collected data was noisy, so we performed the following additional preprocessing beyond [11]:

1. Removing the verb "to be"
2. Removing determiners
3. Removing inflections such as plurals and verb tenses by lemmatization
4. Spelling correction
5. For entities, removing anything that is not a noun (e.g. "red hydrant" to "hydrant")
6. Throwing away all predicates below a frequency threshold

After consolidating any resulting duplicates, we were left with roughly 2,000 unique relationships. In Table 2 we show counts for the most common relationships after preprocessing organized in three groups: spatial, physical and action. Spatial is clearly the most dense type of relationships, but their semantics is often ambiguous.

6.2 Sparsity in Visual Genome

The richness of language and vision is such that, for all its size and complexity, the VG is still far from providing adequate sample coverage of even the simplest grounded spatial relationships.

We analyzed the co-occurrence of entities in images 11 and simple spatial relationships "left" and "right" that are not present in the corpus of annotations, but can be easily extracted from the bounding boxes. We observed a high degree of sparsity.

For instance we discovered that, despite a notable over representation of giraffes, the VG contains only one example of giraffe on the left of a person and eight examples of a person on the left of giraffes, meaning that the same model would need enough capacity to compensate for high imbalance in training data for the two classes.

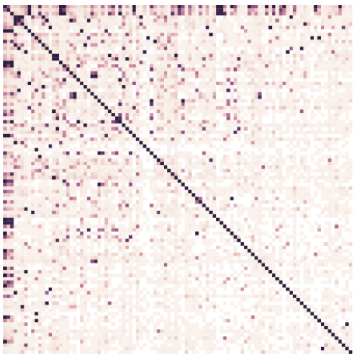

Figure 11: A heatmap of the frequency of top 100 most frequent entities occurring with each other in the same image in VG. Entities are sorted by frequency, and a darker square means higher co-occurrence frequency.

Such imbalances may be inevitable in any human-annotated dataset [15] and we empirically found that these unbalances are present in VG.

6.3 Question generation

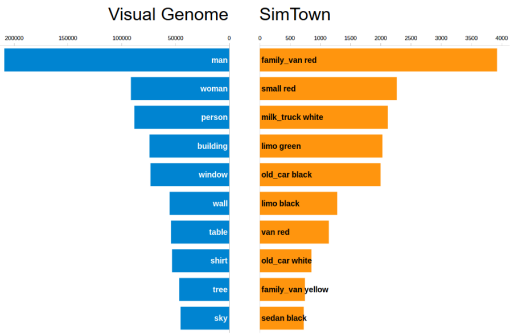

Figure 12: Matching the marginal distributions of subjects and objects in VG and SimTown

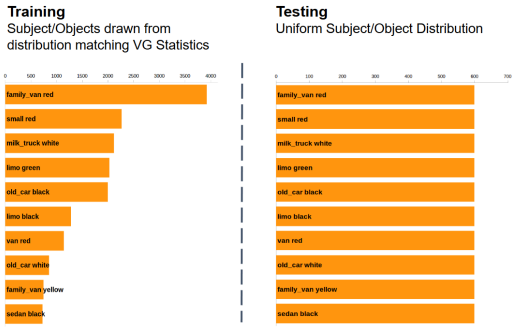

Figure 13: Drawing test sets from the long tail of the training distribution

the probability of generating an utterance (subject s_i, relation r_j, object o_k) is

$$\begin{aligned} P(utterance) &= P(s_i, r_j, o_k) && (2)\\ &= P(s_i)P(r_j)P(o_k) && (3)\\ &= p_{s_i}^{VG} p_{r_j} p_{o_k}^{VG} && (4)\\ &= \frac{1}{N_{rels}} p_{s_i}^{VG} p_{o_k}^{VG} && (5)\\ & && (6) \end{aligned}$$

where

$$p_{s_i}^{VG} = \sum_{m,n} p^{VG}(s_i, r_m, o_n)$$

is the marginal probability of the i-th most frequent subject in the Visual Genome relationship corpus,

$$p_{o_k}^{VG} = \sum_{m,n} p^{VG}(s_m, r_n, o_k)$$

is the marginal probability of the k-th most frequent object in the VG relationship corpus,

and

$$p_{r_j} = \frac{1}{N_{rels}}$$

is the probability of selecting predicate r_j, i.e. we are equally likely to generate any relationship. We chose not to model the covariance between entities because the mapping from objects in SimTown to entities in the Visual Genome is fixed but arbitrary, so the common-sense knowledge encoded in object co-occurrence (e.g. tables are essentially never on top of chairs) is no longer intuitively applicable.

6.4 SimTown sample images

6.5 Experiment details

To control for the effects of clutter in natural images on relationship understanding, we considered two different classes of SimTown datasets: the first class had images with two cars and one relationship annotation, and the second class always had an additional distractor car that was randomly generated and then evaluated to verify that it did not satisfy either the positive or negative relationship associated with the image. All test datasets had 2000 examples.

To preserve spatial information in the visual inputs and make model comparison easier across experiments, two model architectures are trained: a single high-level convolutional network architecture inspired by the VGG ImageNet convolutional network [23] and the all-convolutional networks of [24]. This architecture consisted of 3 layers of 3×3 convolutions with stride 2 followed by batch normalization [5], ReLU activations [12], and dropout regularization [25], finally topped with a multilayer perceptron with ReLU activations with hidden sizes 128, 100, 50, 25, 25, 4. The CNN in the CNN-LSTM model had layers of 20, 50, and 100 filters, and the CNNs in the PBN models always had layers of 32, 64, and 64 filters. The LSTM in the CNN-LSTM model had a single layer with a hidden state size of 128.

We randomly initialized all the weights in all the models and always updated model weights together using the same global learning rate. The PBNs were trained to jointly minimize bounding box regression losses and the verification binary classification loss. All models were trained with stochastic gradient descent with batch size 128 using Adam [10] with learning rate 0.001, and training was stopped after accuracy on a set of $1,000$ held-out validation examples (drawn from the training distributions) did not increase for 10 epochs. All hyperparameters (learning rate, convolutional network size, LSTM hidden state size) were selected for all experiments beforehand by using random search to maximize validation accuracy on the 31 entity, 20k training sample, no distractor dataset.