Data Distributional Properties As Inductive Bias for Systematic Generalization

Supplementary Material

11. Hyper Parameters

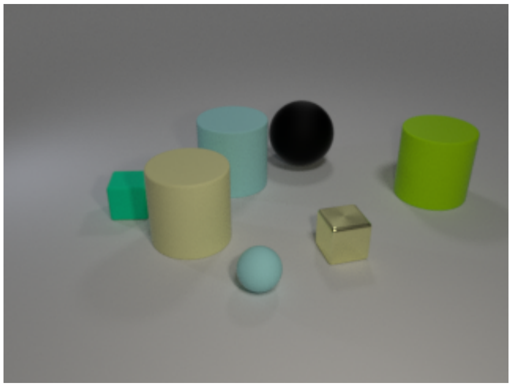
Images were resized to 224 × 224 pixels, and divided into patches of 16 × 16 pixels before being fed to the model. In the text, a masked language modeling (MLM) probability of 0.15 was used to randomly mask text tokens before feeding them to the model.

The model used during our experiments is a Transformer [75] with a hidden layer dimension of 256, with 4 transformer layers and 4 attention heads. Training was carried out using the Adam optimizer [40] with a learning rate of 1×10^{-4}, a batch size of 256 and for 1000 epochs.

12. Dataset Samples

Below we display samples composed of images and their corresponding textual descriptions from the datasets described in Section 3.1:

Sample 1



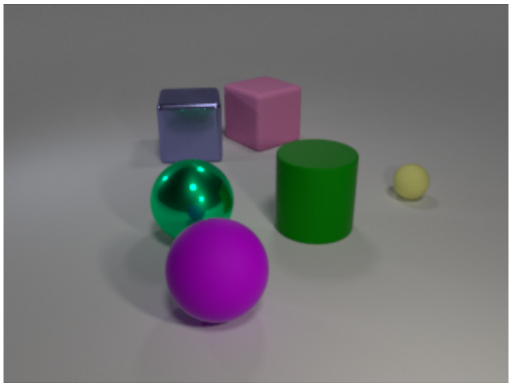
Textual Description

small #00ff80 rubber cube [SEP] large #ffff80 rubber cylinder [SEP] large #80ffff rubber cylinder [SEP] small #ffff80 metal cube [SEP] large #000000 rubber sphere [SEP] small #80ffff rubber sphere [SEP] large #80ff00 rubber cylinder

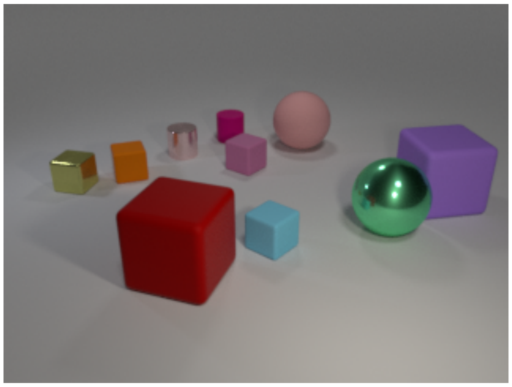
Sample 2

Textual Description

large #005500 rubber cylinder [SEP] large #5555aa metal cube [SEP] large #ff55aa rubber cube [SEP] small #ffff55 rubber sphere [SEP] large #aa00ff rubber sphere [SEP] large #00ff55 metal sphere



Sample 3



Textual Description

small #bf0040 rubber cylinder [SEP] large #40ff80 metal sphere [SEP] large #8040ff rubber cube [SEP] large #800000 rubber cube [SEP] small #bf4080 rubber cube [SEP] small #ffbfbf metal cylinder [SEP] small #ff4000 rubber cube [SEP] small #40bfff rubber cube [SEP] small #bfbf40 metal cube [SEP] large #ff8080 rubber sphere

13. Correlation between P-Score and OOD Performance

In Table 1 we show the Pearson correlation between different p-scores obtained for a model and their OOD performance on the shape task. Parallelism in the representations seems to strongly correlate with better OOD performance, suggesting that parallelism may be the property that is inducing SG in the models.

METRIC	AVG	DIV	BURST	LI
CORR	0.73	0.76	0.69	0.73
P-VALUE	$1.88e{-9}$	$7.17e{-4}$	$5.5e{-3}$	$3.45e{-5}$

Table 1. Pearson Correlation between P-Scores and OOD performance on the *shape* task broken down per training data property. DIV: Diversity, BURST: Burstiness, LI: Latent Intervention.

14. Detailed Experimental Results

In tables 2, 3, 4, 5, 6, and 7, we exposed the values obtained that were used to generate most of the figures in the work above. Table 2 shows the data used for Figure 2. Table 3 for Figure 3. Table 4 for Figure 4. Table 5 for Figure 6. Table 6 for Figure 5. And, Table 7 for Figure 7.

	Train	Test-ID	Test-OOD
Color	94.3 ± 0.0%	94.1 ± 0.1%	50.5 ± 1.1%
Shapes	100.0 ± 0.0%	99.6 ± 0.0%	0.6 ± 0.1%
Materials	98.0 ± 0.0%	96.9 ± 0.0%	87.8 ± 0.9%
Size	98.1 ± 0.0%	98.0 ± 0.0%	91.2 ± 1.0%

Table 2. Table displaying values to form Figure 2. Accuracy for predicting different properties for different data splits for the baseline model (8 colors). ID performance for all tasks remains high, however OOD performance plummets for shape and color, suggesting the model is learning combinations of shape-color as features, instead of achieving SG. Unexpectedly, material and size also show a drop in OOD performance, even though the model has been exposed to all combinations of these attributes in training.

	Color		Shape	
Train Fract.	Test-ID	Test-OOD	Test-ID	Test-OOD
8 Colors				
1/8	91.5 ± 0.1%	47.4 ± 0.4%	97.1 ± 0.1%	2.1 ± 0.1%
1/4	93.1 ± 0.1%	47.4 ± 0.5%	99.0 ± 0.0%	1.2 ± 0.1%
1/2	93.9 ± 0.0%	47.3 ± 0.6%	99.2 ± 0.0%	0.6 ± 0.1%
1/1	94.1 ± 0.1%	50.5 ± 1.1%	99.6 ± 0.0%	0.6 ± 0.1%
216 Colors				
1/8	22.8 ± 2.4%	5.0 ± 0.9%	45.1 ± 0.6%	43.3 ± 0.5%
1/4	79.6 ± 0.6%	60.4 ± 1.3%	89.1 ± 0.5%	81.0 ± 0.8%
1/2	86.1 ± 0.0%	74.1 ± 0.4%	94.4 ± 0.1%	88.4 ± 0.2%
1/1	88.1 ± 0.1%	77.9 ± 0.4%	96.3 ± 0.0%	90.0 ± 0.4%

Table 3. Table displaying values to form Figure 3. Accuracy for the *shape* task for different amounts of training data for the baseline model (8 colors) vs model trained on 216 colors. (a) Increasing dataset size does not increase OOD performance but rather degrades it slightly. (b) With increased *diversity* much stronger OOD generalization is achieved, with models trained on only a quarter of the data severely outperforming the 8-color baseline.

Num. Colors	Color		Shape		Size		Material	
	Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD
8	94.1 ± 0.1%	50.5 ± 1.1%	99.6 ± 0.0%	0.6 ± 0.1%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
27	91.7 ± 0.0%	67.5 ± 0.7%	96.9 ± 0.1%	1.5 ± 0.1%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
64	90.8 ± 0.1%	73.0 ± 0.5%	96.9 ± 0.0%	48.5 ± 0.6%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
125	89.0 ± 0.1%	67.0 ± 0.3%	96.1 ± 0.1%	81.8 ± 0.6%	99.8 ± 0.0%	97.8 ± 0.0%	99.2 ± 0.0%	96.9 ± 0.0%
216	88.1 ± 0.1%	77.9 ± 0.4%	96.3 ± 0.0%	90.0 ± 0.4%	99.8 ± 0.0%	97.7 ± 0.0%	99.3 ± 0.0%	97.2 ± 0.0%

Table 4. Table displaying values to form Figure 4. Accuracy versus the number of colors in $\mathcal{D}_{train}$ for $\mathcal{D}_{test-ID}$ and $\mathcal{D}_{test-OOD}$ for all tasks. Performance for the *shape* task increases drastically for the OOD split as we increase colors, increasing 86% in absolute terms over the 8-color baseline. Moreover, performance in the color task also tends to increase in the OOD split, while ID only suffers slightly, even though the task becomes significantly harder. Remarkably, the *material* and *size* task rapidly increase their $\mathcal{D}_{test-OOD}$ performance as color increases.

Num. Colors	P Bursty	Color		Shape		Size		Material	
		Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD
8	0.0	94.1 ± 0.1%	50.5 ± 1.1%	99.6 ± 0.0%	0.6 ± 0.1%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.5	94.2 ± 0.0%	48.9 ± 0.9%	99.6 ± 0.0%	0.9 ± 0.2%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	1.0	93.8 ± 0.0%	47.3 ± 0.1%	99.4 ± 0.0%	0.5 ± 0.0%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
27	0.0	91.7 ± 0.0%	67.5 ± 0.7%	96.9 ± 0.1%	1.5 ± 0.1%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.5	91.7 ± 0.0%	71.9 ± 0.2%	97.1 ± 0.0%	1.8 ± 0.3%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	1.0	90.3 ± 0.0%	65.2 ± 0.8%	96.9 ± 0.0%	4.7 ± 0.4%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
64	0.0	90.8 ± 0.1%	73.0 ± 0.5%	96.9 ± 0.0%	48.5 ± 0.6%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.5	89.8 ± 0.5%	73.8 ± 1.0%	97.0 ± 0.1%	60.5 ± 2.0%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	1.0	83.5 ± 0.2%	58.7 ± 1.0%	96.9 ± 0.1%	63.3 ± 0.9%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
125	0.0	89.0 ± 0.1%	67.0 ± 0.3%	96.1 ± 0.1%	81.8 ± 0.6%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.5	88.9 ± 0.1%	68.5 ± 0.3%	96.8 ± 0.0%	80.6 ± 1.2%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	1.0	80.7 ± 0.2%	64.6 ± 0.3%	96.8 ± 0.0%	79.4 ± 0.5%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
216	0.0	88.1 ± 0.1%	77.9 ± 0.4%	96.3 ± 0.0%	90.0 ± 0.4%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.5	87.3 ± 0.0%	77.7 ± 0.2%	96.7 ± 0.0%	91.4 ± 0.2%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	1.0	71.9 ± 0.2%	57.3 ± 0.4%	96.7 ± 0.0%	92.4 ± 0.3%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%

Table 5. Table displaying values to form Figure 5. Accuracy fo all tasks in test-ID and test-OOD for different levels of *burstiness* over *color* for various numbers of colors. Limiting the number of colors available for each image during training allows the model to gain up to 14.8% more accuracy over the baseline. The *color* task, however, suffers up to 14.3% decline as the *color* task becomes easier to memorize.

		Color		Shape		Size		Material	
Num. Colors	Jitter	Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD
8	0.00	94.1 ± 0.1%	50.5 ± 1.1%	99.6 ± 0.0%	0.6 ± 0.1%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.05	94.1 ± 0.1%	46.6 ± 0.4%	99.6 ± 0.0%	0.4 ± 0.0%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	0.10	94.1 ± 0.1%	48.0 ± 0.7%	99.6 ± 0.0%	0.9 ± 0.2%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
	0.50	93.9 ± 0.0%	47.1 ± 0.4%	99.7 ± 0.0%	0.3 ± 0.1%	99.8 ± 0.0%	97.8 ± 0.0%	99.2 ± 0.0%	96.9 ± 0.0%
27	0.00	91.7 ± 0.0%	67.5 ± 0.7%	96.9 ± 0.1%	1.5 ± 0.1%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.05	91.9 ± 0.0%	69.4 ± 1.5%	97.4 ± 0.0%	1.4 ± 0.2%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	0.10	91.8 ± 0.1%	68.4 ± 1.0%	97.3 ± 0.1%	1.7 ± 0.2%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
	0.50	91.9 ± 0.0%	66.1 ± 2.0%	97.2 ± 0.0%	1.1 ± 0.2%	99.8 ± 0.0%	97.8 ± 0.0%	99.2 ± 0.0%	96.9 ± 0.0%
64	0.00	90.8 ± 0.1%	73.0 ± 0.5%	96.9 ± 0.0%	48.5 ± 0.6%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.05	90.9 ± 0.0%	68.5 ± 0.6%	97.1 ± 0.0%	57.1 ± 0.1%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	0.10	91.0 ± 0.0%	67.1 ± 0.3%	97.3 ± 0.0%	62.3 ± 3.0%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
	0.50	91.1 ± 0.0%	68.6 ± 0.6%	97.8 ± 0.0%	63.8 ± 0.9%	99.8 ± 0.0%	97.8 ± 0.0%	99.2 ± 0.0%	96.9 ± 0.0%
125	0.00	89.0 ± 0.1%	67.0 ± 0.3%	96.1 ± 0.1%	81.8 ± 0.6%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.05	89.8 ± 0.0%	71.1 ± 0.4%	96.8 ± 0.1%	85.0 ± 0.5%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	0.10	89.7 ± 0.1%	68.4 ± 0.5%	96.7 ± 0.1%	79.2 ± 0.7%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
	0.50	89.8 ± 0.1%	70.6 ± 1.1%	96.9 ± 0.0%	85.8 ± 0.9%	99.8 ± 0.0%	97.8 ± 0.0%	99.2 ± 0.0%	96.9 ± 0.0%
216	0.00	88.1 ± 0.1%	77.9 ± 0.4%	96.3 ± 0.0%	90.0 ± 0.4%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	0.05	88.4 ± 0.0%	79.8 ± 0.3%	96.8 ± 0.1%	92.2 ± 0.1%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	0.10	88.0 ± 0.1%	75.6 ± 0.2%	96.4 ± 0.1%	91.0 ± 0.3%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
	0.50	88.4 ± 0.1%	78.7 ± 0.7%	96.7 ± 0.1%	92.0 ± 0.1%	99.8 ± 0.0%	97.8 ± 0.0%	99.2 ± 0.0%	96.9 ± 0.0%

Table 6. Table displaying values to form Figure 6. Accuracy after applying latent intervention for all tasks in test-ID and test-OOD for different levels of latent intervention of the *color* latent attribute for various numbers of colors. Altering the color hue randomly during training allows the model to gain up to 15% more OOD accuracy over the baseline.

		Color		Shape		Size		Material	
Num. Colors	Hidden Size	Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD	Test-ID	Test-OOD
8	32	79.3 ± 0.1%	4.6 ± 0.6%	90.8 ± 0.8%	0.0 ± 0.0%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	64	87.0 ± 1.6%	18.1 ± 1.9%	98.1 ± 0.4%	0.0 ± 0.0%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	128	93.8 ± 0.1%	43.9 ± 0.7%	99.5 ± 0.0%	0.2 ± 0.0%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%
	256	94.1 ± 0.1%	50.5 ± 1.1%	99.6 ± 0.0%	0.6 ± 0.1%	99.8 ± 0.0%	97.8 ± 0.0%	99.2 ± 0.0%	96.9 ± 0.0%
216	256	88.1 ± 0.1%	77.9 ± 0.4%	96.3 ± 0.0%	90.0 ± 0.4%	98.0 ± 0.0%	91.2 ± 1.0%	96.9 ± 0.0%	87.8 ± 0.9%
	512	88.8 ± 0.0%	81.6 ± 0.2%	97.3 ± 0.0%	93.5 ± 0.1%	99.5 ± 0.0%	97.6 ± 0.0%	98.6 ± 0.0%	96.7 ± 0.1%
	1024	89.1 ± 0.1%	82.2 ± 0.1%	97.6 ± 0.0%	92.4 ± 0.2%	99.7 ± 0.0%	97.6 ± 0.0%	99.2 ± 0.0%	96.8 ± 0.0%

Table 7. Table displaying values to form Figure 7. ID and OOD accuracy for models trained with different values for their hidden dimensions on a training set with 8 and 216 colors for all tasks.