A Adjective Frequencies

Figure 4 shows a histograme of the most frequent adjectives in the captions of the COCO dataset.

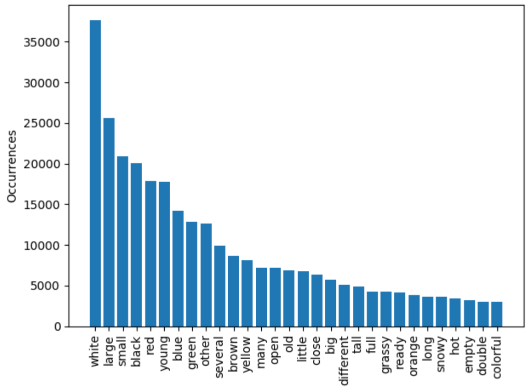

Figure 4: Histogram of the adjectives used in COCO.

B Concept Pairs Statistics

Table 5 shows the number of images for which at least one reference caption includes the respective concept pair. The two numbers indicate scores for the COCO training set (which is also used for training, by holding out exactly this set of images) and the COCO validation set (which is used for evaluation).

C Dataset splits

To increase the efficiency of training and evaluation, we create training sets in which we simultaneously hold out multiple pairs. We ensure that no more than 5% of the training data is removed from the original training set, and that we do not remove pairs with overlapping nouns, adjectives or verbs within the same training set.

Based on these constraints, we create four sets of training and evaluation splits. Each set contains a held out pair for a color modifier on an animate and inanimate object, a size modifier on an inanimate and inanimate object and a transitive and an intransitive verb for animate objects. For each of these four splits, we train a model on the respective training data and calculate the recall for each held out pair on the respective evaluation data.

Further, we calculate average recall scores for various groups of conceptually similar held out pairs and an average over all recall scores as a single measure indicating the compositional generalization performance of a model.

	Training Set	Validation Set
`black bird`	205	122
`small dog`	681	316
`white boat`	373	196
`big truck`	417	191
`eat horse`	212	106
`stand child`	1288	577
`white horse`	264	151
`big cat`	184	103
`blue bus`	276	143
`small table`	261	134
`hold child`	1328	664
`stand bird`	532	260
`brown dog`	613	291
`small cat`	252	149
`white truck`	262	121
`big plane`	967	357
`ride woman`	595	300
`fly bird`	245	132
`black cat`	840	448
`big bird`	215	123
`red bus`	566	232
`small plane`	481	158
`eat man`	555	250
`lie woman`	301	144

Table 5: Number of occurrences of concept pairs in the COCO training and validation set. The full training set size is 82,783 images, the validation set consists of 40,504 images.

Table 6 lists the held out word pairs and their distribution into four different datasets. We did not include inanimate verb–noun pairs because there were not enough instances in the validation set.

D Synonyms

Table 7 shows the synonyms we defined for our selected adjectives and verbs. For the noun synonyms, refer to Lu et al. (2018, Appendix)

E Training BUTR

In this section we describe the hyperparameters and training details of BUTR. The parameters have been chosen in accordance with the BUTD and VSE++ models and not further tuned. BUTR is trained with a 1024D visual-semantic embedding space (J), a 1000D language encoding LSTM (L), a 1000D language generation LSTM (G), a vocabulary of 10000 types (V), 300D word embeddings

	Held out pairs	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{eval}}$
1	black cat, big bird, red bus, small plane, eat man, lie woman	79,825	1,355
2	brown dog, small cat, white truck, big plane, ride woman, fly bird	79,849	1,350
3	white horse, big cat, blue bus, small table, hold child, stand bird	79,938	1,455
4	black bird, small dog, white boat, big truck, eat horse, stand child	79,607	1,508

Table 6: The held out word pairs in each dataset split. Training and evaluation set sizes are in number of images; each image is associated with five captions. The full training set size is 82,783 images.

Word	Synonyms
big	large, tall, huge, wide, great, broad, enormous, expansive, extensive, giant, gigantic, massive, vast
small	little, narrow, short, tinier, tiny, thin, compact, mini, petite, skinny
red	dark-red, light-red
brown	brownish, dark-brown, light-brown
blue	blueish, light-blue, dark-blue
black	-
white	-
eat	chew, bite, graze
lie	lay
hold	carry
ride	-
fly	-
stand	-

Table 7: The adjective and verb synonyms used to select word pairs for the experiments in this paper.

(E), 2048D image region feature vectors, a 512D attention model dimension, and inference is performed using beam search with a 100 hypotheses (B). BUTR is trained using pre-computed bottom-up image features from 36 regions obtained using the bottom-up encoder defined in Anderson et al. (2016). The caption generation component is trained with teacher forcing and a maximum caption length of 20 in batches of 100 with the Adam optimizer (Kingma and Ba, 2014) using an initial learning rate of 1e-4. The gradients are clipped when they exceed 10.0. For the GradNorm optimizer, we also use Adam, but with an initial learning rate of 0.01. We set the asymmetry to 2.5. BUTR is trained for at most 30 epochs, and early stop when the validation set BLEU score does not increase for five consecutive epochs.

F Describing Sizes

To support the claim that the bounding box sizes do not necessarily relate to the actual sizes of the objects as they are described, we perform a correlation analysis. We make use of the fact that there is bounding box annotations for objects in the COCO dataset. We identify each noun concept that was also used in combination with size modifiers in the held out concept pairs (cf. Table 1: cat, plane, table, dog, bird, and truck. For each of these concepts, we consider all images that contain at least one instance of the object as annotated in the COCO dataset. Given one of these images, we regard only the size of the area of the biggest bounding box[9] belonging to an object of that kind. Then, we look at the reference captions belonging to the respective image and look for matching concept pairs[10]. To test whether the bounding box sizes relate to the described sizes of the objects, we perform a unpaired t-test comparing the box sizes for objects described as small and objects described as big.

Table 8 shows the average bounding box size for the set of concept pairs. Further, the last column shows the resulting p-values from the t-tests. The differences in box sizes for small vs. big objects are never significant, except for the case of table ($p \approx 0.007$). However, in this case the box sizes are on average bigger if the table is described as small. We conclude that the bounding box sizes of objects in the COCO dataset do not relate to the described sizes in the respective captions.

G Describing Actions

We analyze the dataset and calculate statistics on the occurrence of objects in connection with the concept pairs that include transitive and intransitive verbs. We use StanfordNLP for detecting the

[9] We assume that the biggest object of a category in the image is also the most salient and thus most likely the one that was described.

[10] We disregard all images with contradicting descriptions (i.e. different annotators describe the object as small and big) and images where the size of the concept is not described at all.

Concept	Average bounding box size (in pixels)	Number of samples	p-Value
small cat	42,920.6 ± 38,952.2	628	0.64
big cat	44,057.4 ± 41,979.9	495	
small plane	33,718.8 ± 30,481.2	569	0.77
big plane	33,263.1 ± 31,722.9	1,408	
small dog	36,939.5 ± 41,073.3	1,109	0.94
big dog	37,098.3 ± 40,088.6	718	
small table	80,762.0 ± 89,751.0	1,860	0.007
big table	72,958.0 ± 91,340.0	2,101	
small bird	15,063.0 ± 19,487.6	774	0.77
big bird	14,707.8 ± 27,008.7	789	
small truck	30,014.0 ± 49,121.4	531	0.21
big truck	32,918.2 ± 46,379.8	1,945	

Table 8: Comparison of bounding box sizes for different concept pairs describing sizes of objects. The last column indicates the resulting p-value from an unpaired t-test between the data of the two respective rows.

Concept Pair	with Object	including ”obl”
hold child	96%	99%
ride woman	81%	97%
eat man	87%	97%
stand child	26%	92%
stand bird	3%	98%
fly bird	7%	89%
lie woman	24%	96%

Table 9: Percentage of captions where a direct or indirect object is connected to the noun of the concept pair. In the last column, additional arguments (”obl”) are also counted as objects.

objects. The examined concept pairs for transitive verbs are `hold child`, `ride woman`, `eat man` and for intransitive verbs `stand child`, `stand bird`, `fly bird`, and `lie woman`.[11]

The results are presented in Table 9. In fact, phrases using transitive verbs contain objects 88% of the time and phrases using intransitive verbs only 15% of the time. If we include additional arguments (marked as oblique ”obl”) in our definition of objects, the percentage in the transitive verb case rises to 98%, and in the intransitive case to 93%. An unpaired t-test shows that this difference is still significant ($p < 10^{-38}$).

The performed analysis supports the hypothesis that the models perform better for actions described with transitive verbs because of additional clues coming from the object.

H Detailed Results

Table 10 presents the Recall@5 generalization performance for each held out pair.

[11] We exclude the pair `eat horse` from the analysis, because we defined ”graze” as a synonym for ”eat” (cf. Table 7 which is an intransitive verb. We find that this is quite often used and thus would decrease the validity of the statistics

	SAT	**BUTD**	**BUTR**	**+RR**	**FULL**
black bird	7.4	1.6	4.1	**9.8**	25.4
small dog	0	**0.3**	0	**0.3**	13.0
white boat	1.5	5.1	4.6	**8.2**	17.3
big truck	0	0	0	**0.5**	35.1
eat horse	0	19.8	7.5	**36.8**	41.5
stand child	0.7	3.6	3.1	**14.0**	24.4
white horse	4.0	10.6	9.9	**13.9**	48.3
big cat	0	0	0	0	0
blue bus	15.4	6.3	22.4	**28.0**	40.6
small table	0	0	0	0	0.7
hold child	3.2	5.9	3.2	**11.6**	33.7
stand bird	1.2	6.9	5.8	**11.2**	41.2
brown dog	0.3	1.4	3.8	**9.3**	29.9
small cat	0	0	**1.3**	**1.3**	0.7
white truck	8.3	8.3	8.3	**19.0**	31.4
big plane	0	0	0.8	**2.5**	58.3
ride woman	0	10.7	3.7	**15.3**	46.0
fly bird	6.1	19.7	21.2	**25.0**	52.3
black cat	3.1	7.8	7.8	**22.3**	67.2
big bird	0	1.6	0	**4.1**	9.8
red bus	16.8	24.1	29.7	**48.7**	65.5
small plane	0	0	0	0	39.2
eat man	3.2	10	13.6	**17.6**	37.2
lie woman	0.7	11.1	4.2	**17.4**	40.3

Table 10: Recall@5 for each of the held out concept pairs. RR stands for re-ranking after decoding. The **bold face** results denote the best model performance when trained with paradigmatic gaps.