Combine to Describe: Evaluating Compositional Generalization in Image Captioning

Georgios Pantazopoulos
Heriot-Watt University
Edinburgh, Scotland
gmp2000@hw.ac.uk

Alessandro Suglia
Heriot-Watt University
Edinburgh, Scotland
a.suglia@hw.ac.uk

Arash Eshghi
Heriot-Watt University
Edinburgh, Scotland
a.eshghi@hw.ac.uk

Abstract

Compositionality – the ability to combine simpler concepts to understand & generate arbitrarily more complex conceptual structures – has long been thought to be the cornerstone of human language capacity. With the recent, notable success of neural models in various NLP tasks, attention has now naturally turned to the compositional capacity of these models. In this paper, we study the compositional generalization properties of image captioning models. We perform a set of experiments under controlled conditions using model and data ablations, each designed to benchmark a particular facet of compositional generalization: *systematicity* is the ability of a model to create novel combinations of concepts out of those observed during training, *productivity* is here operationalised as the capacity of a model to extend its predictions beyond the length distribution it has observed during training, and *substitutivity* is concerned with the robustness of the model against synonym substitutions. While previous work has focused primarily on systematicity, here we provide a more in-depth analysis of the strengths and weaknesses of state of the art captioning models. Our findings demonstrate that the models we study here do not compositionally generalize in terms of systematicity and productivity, however, they are robust to some degree to synonym substitutions[1].

1 Introduction

Deep neural networks have undoubtedly become the standard option for many Natural Language Processing (NLP) tasks with tangible results across a variety of tasks (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020; Liu et al., 2019). Despite their success, neural networks are regularly criticized from a growing body of research for their limited capacity to generalize beyond the distribution of the data on which they were trained. A frequent topic of discussion is *compositionality* of meaning. Humans can understand or generate novel and more complex conceptual structures or sentences out of simpler constituent representations, without needing to encounter any instances of these more complex structures. On the other hand, to what extent different neural models exhibit compositional behavior remains an open problem (Fodor and Pylyshyn, 1988; Smolensky, 1990; Hupkes et al., 2020; Baroni, 2020).

Previous work studying compositionality has primarily focused on artificially created datasets, where compositional rules can be isolated from other natural language phenomena (Baroni, 2020). In this setting, the majority of prior work has largely focused on systematicity under the prism of a downstream task. While these approaches have provided valuable insights, compositionality is multifaceted and a single test can yield misleading findings regarding compositional generalization.

In this paper, we explore compositionality from the perspective of image captioning as a grounded natural language task and propose an evaluation framework with multiple dimensions of compositionality for captioning models. In particular, we adapt the independent and task-agnostic compositionality tests from Hupkes et al. (2020) to the task of image captioning using data and model ablations. Each test is designed to quantify the behavior of a model along a specific dimension of compositionality. In particular, we evaluate three facets of compositionality: (1) *systematicity* (Fodor and Pylyshyn, 1988; Fodor and Lepore, 2002): the ability to generalize to unseen combinations of concepts learned in isolation during training; (2) *productivity*: the capacity to extend predictions beyond the observations; and (3) *substitutivity*: the robustness of predictions under synonym substitution. Previous approaches investigating compositionality in image captioning have focused primarily on systematicity (Atzmon et al., 2016; Nikolaus et al.,

[1] Code & data are available here.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 115 - 131
May 22-27, 2022 ©2022 Association for Computational Linguistics

2019; Bugliarello and Elliott, 2021). Our work thus constitutes a more in-depth analysis on the compositional capabilities of captioning models.

Our findings regarding systematicity indicate that the standard fine-tuning approach using reinforcement learning provides gains in word-overlap metrics but hinders systematic generalization. In productivity, we demonstrate that models struggle to extend the length of their prediction beyond the training distribution. Finally, with substitutivity, we demonstrate that state of the art captioning models we study here are robust against substitutions of fine-grained with more high-level synonyms.

2 Related Work

In mathematical logic, the *principle of compositionality* declares that the meaning of an expression can be derived from the meanings of its constituent expressions (Frege, 1950). From the perspective of natural language, if all lexical/word meaning is abstracted out from a sentence, then what remains are the rules of composition. Implications of the principle influence research to this day with longstanding debates regarding compositional properties of vector space and neural models.

Compositionality in Neural Language Processing Initial approaches on distributional, vector-space semantics use tensors as word and phrase meaning representations, and has studied various tensor operations for composition (Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Coecke et al., 2010; Sadrzadeh and Grefenstette, 2011; Purver et al., 2021). In all this work, the compositional operations are fixed in advance based on some linguistic theory. In contrast, neural models learn to encode meaning: compositional operations are neither fixed during processing, nor given in advance. To encourage compositionality of neural models, prior work clusters around data augmentation (Akyürek et al., 2020; Qiu et al., 2021), loss functions that encode different inductive biases (Yin et al., 2021; Jiang and Bansal, 2021), as well as meta-learning (Conklin et al., 2021).

Benchmarking compositionality Compositionality is often measured as systematic generalization in different tasks including: in navigation environments (Lake and Baroni, 2018), where the objective is to translate commands into sequences of actions; or in question-answering, (Sinha et al., 2019; Keysers et al., 2019; Kim and Linzen, 2020), where to answer a question the model needs to infer underlying relationships between entities. Additional benchmarks include evaluating arithmetic expressions (Veldhoen et al., 2016; Saxton et al., 2019), and logical entailment (Bowman et al., 2015; Mul and Zuidema, 2019).

Compositionality in Visually Grounded Natural Language Compositionality has also been studied from the perspective of visually grounded natural language. Previous work on visual question answering (VQA) measures generalization to novel question-answer pairs on natural (Agrawal et al., 2017; Whitehead et al., 2021), and synthetic datasets (Bahdanau et al., 2018; Johnson et al., 2017). Similarly, Suglia et al. (2020), proposes an evaluation framework that accounts for a model's systematic generalization capacity coupled with task performance in the context of visual guessing games. More closely related to this paper, prior work has examined compositionality for image captioning (Atzmon et al., 2016; Nikolaus et al., 2019; Bugliarello and Elliott, 2021). However, the aforementioned works mainly focus on systematicity alone and thus provide valuable, but limited insights on compositional properties.

Some prior work has studied compositionality along different prisms. Ruis et al. (2020) examines compositionality under multiple dimensions extending the work of Lake and Baroni (2018) by grounding language to grid world environments. Hupkes et al. (2020) provides a multifaceted view on compositional properties of neural models under a set of task-agnostic tests instantiated with an artificial translation task. The distilled conclusion of this work is that the performance on a single downstream task is not a representative indicator of compositional awareness, even if this task is designed to be highly compositional. This paper can be viewed as an extension of the latter line of work, where we adapt the more fine-grained compositionality tests to the visually grounded image captioning task.

3 Testing Compositionality in Image Captioning

In this section, we describe the proposed tests for evaluating compositionality in image captioning. Figure 1 illustrates examples from each test. We adopt a subset of task-agnostic tests proposed by Hupkes et al. (2020). Our suite consists of three tests: systematicity, productivity, and substitutivity.

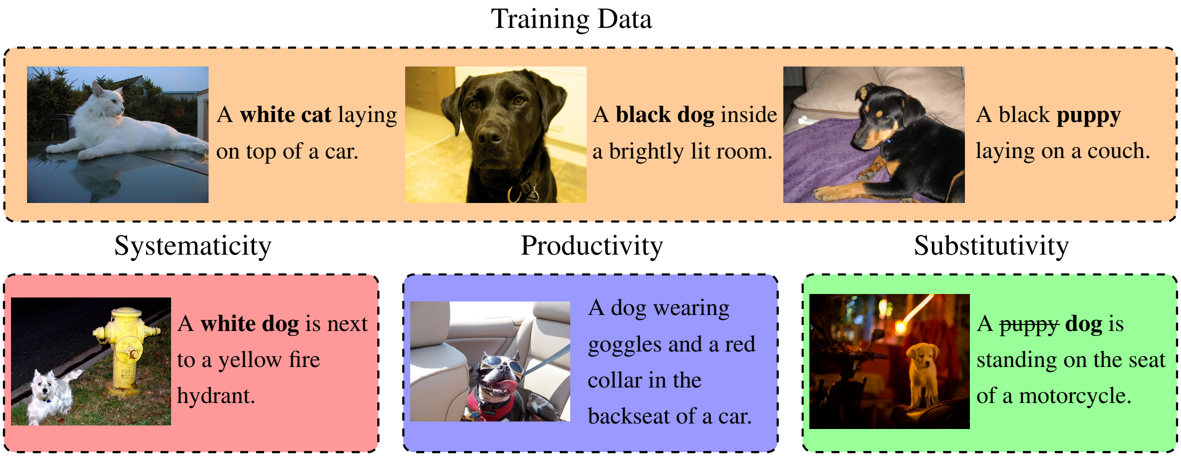

Training Data
A **white cat** laying on top of a car.
A **black dog** inside a brightly lit room.
A black **puppy** laying on a couch.
Systematicity
Productivity
Substitutivity
A **white dog** is next to a yellow fire hydrant.
A dog wearing goggles and a red collar in the backseat of a car.
A ~~puppy~~ **dog** is standing on the seat of a motorcycle.

Figure 1: Illustration of different compositionality tests. In the systematicity test, we evaluate the ability to combine concepts (white cat, black dog) to form novel output (white dog). With productivity we focus on the conditions where a model can produce output extending beyond the observable samples. In the substitutivity test, we investigate if models are robust to synonym substitutions. Note that the training data is different across each compositional test.

For each test we define custom training and evaluation splits of the MSCOCO dataset (Lin et al., 2014). Appendix A contains additional details concerning dataset splits, and we will release this data in the public domain.

Systematicity The first test asserts the model's ability to combine known concepts into new expressions. If somebody can understand the meaning of a 'black dog' and a 'white cat', then they can understand the meaning of a 'white dog' (Szabó, 2012). Consequently, a model should be able to describe a white dog even though it has only observed pairs of black dogs and white cats during training.

To probe for systematic behavior, we consider pairs of concepts where their combination is observed during testing and independently during training. In the above example, the pairs 'black dog' and 'white cat' belong to the training while the pair 'white dog' is assigned to the evaluation split. Following Nikolaus et al. (2019), we adopt systematicity splits with pairs of adjectives and nouns, as well as verbs and nouns. For comparative analysis we use a second evaluation set where the constituents of the pairs are observed separately.

Productivity Natural languages are said to be productive in the sense that the speakers of a language are able to understand & generate a theoretically infinite set of expressions or sentences. While there is broad agreement that much of this productivity is buttressed by systematicity, there are also exceptional cases of non-systematic, or partial productivity (Baroni, 2020).

In this paper, we operationalise the broad concept of productivity in two very specific ways. First, we take the productivity of a model to be its ability to generate captions beyond the length it has observed during training (Graves et al., 2014). We tokenize each caption and compute the average caption length for each image. We assign the images at the tail of the histogram of the average caption length to the evaluation set. From the remaining pool of images we sample a second equally-sized evaluation set for comparative analysis. The remaining images are used for training. Second, we assume a model exhibits productive behavior if it can describe significantly denser, more complex images than it has observed during training. We use the number of ground truth bounding boxes from MSCOCO as an indicator of the number of objects present in a scene, and as a measure of their density; and use this measure to create controlled training and evaluation splits.

Substitutivity Substitutivity states that the meaning of a complex expression is not altered after replacement of one of its constituents with another constituent that has the same meaning (Pagin, 2003). Therefore, if a model is compositional then replacing an expression with its synonym should not affect the structure nor the meaning of the whole expression. In the above example a model should be able to infer that the word 'puppy' and 'dog' are synonyms, thus the substitution should preserve the meaning and structure of the caption.

We use a subset of the synonyms of the 80 MSCOCO categories defined by Lu et al. (2018).

We consider substitutions between the original object category and the corresponding retrieved synonym. For each object category and its synonyms, we select pairs that make valid substitutions given the visual context by manually inspecting ground truth captions containing each constituent word. We further exclude ambiguous words and divide object categories to ensure that the substitutions we make are always valid. For instance, we divide the 'person' category into 'man', 'woman' and 'child'.

4 Experiments and results

Model We use $\mathcal{M}^2$-transformer (Cornia et al., 2020), and adopt the configuration with the best reported results. Following standard protocol (Anderson et al., 2018; Lu et al., 2018; Cornia et al., 2020), the training scheme in all experiments consists of two phases: cross-entropy (XE) and CIDEr optimization. For XE, we apply teacher forcing where the model is trained to predict the next token given the previous ground truth tokens. We adopt Self Critical Sequence Training (SCST) (Rennie et al., 2017), as the reinforcement learning paradigm for CIDEr optimization. The reward function is the CIDEr score obtained with sampled sentences using beam search. For both phases, we applied the same training hyperparameters as in Cornia et al. (2020). The training phases are performed sequentially. We start by optimizing XE and then fine-tune the model using SCST. We used early stopping to terminate a training phase, whenever the CIDEr score on the validation set did not improve for 5 consecutive epochs.

Evaluation We evaluate compositional generalization using standard metrics in image captioning: BLEU (B1, B4, Papineni et al., 2002), METEOR (M, Denkowski and Lavie, 2014), ROUGE (R, Lin, 2004), and CIDEr (C, Vedantam et al., 2015. We also quantify semantic similarity using the multi-reference BERTSCORE (Yi et al., 2020). For the case of systematicity, we follow previous approaches (Nikolaus et al., 2019; Bugliarello and Elliott, 2021), and additionally report Recall@K of the pair of interest over the K generated captions using beam search ($K = 1 \ldots 5$).

4.1 Systematicity

In the first set of experiments, we investigate whether or not the model can combine known concepts disjointly observed during training. We adopt a subset of the pairs of adjectives and nouns, verbs and nouns defined by Nikolaus et al. (2019), and modify the proposed train, validation, and test sets. The examined pairs are presented in Table 7.

		B1	B4	M	R	C	BS
V	$\mathcal{M}^2$	75.71	37.01	27.81	58.22	106.25	45.07
	$\mathcal{M}^2_{SCST}$	78.68	39.37	28.93	59.51	116.81	46.15
TNC	$\mathcal{M}^2$	75.22	35.94	27.35	56.42	115.95	43.78
	$\mathcal{M}^2_{SCST}$	80.36	39.14	28.59	58.41	130.16	45.20
TC	$\mathcal{M}^2$	75.83	36.08	27.56	57.79	105.90	44.83
	$\mathcal{M}^2_{SCST}$	79.02	38.66	28.56	59.36	116.11	45.80

Table 1: Results on systematicity split in validation (V), Test no Comb (TNC), and Test Comb (TC).

With these pairs we test the model under two different conditions: Test no Comb (TNC) consists of images where the constituents of the pairs are not observed in the same image; and Test Comb (TC) is the test set defined in Nikolaus et al. (2019). For TNC, we sampled random images from the proposed train set ensuring that both test sets have the same number of images. Finally, we used the same validation set as the proposed split. Notably, the validation set consists of images where at least one of the captions contains the concept of interest. This means that while the model is not directly exposed to the combination of the concepts, it is tuned by optimizing the evaluation metrics on a set that contains these combinations.

Table 1 shows the performance of both models in terms of standard captioning metrics. In particular, SCST improves the performance of the model in terms of word similarity metrics, but also in terms of semantic equivalence as shown by BERTSCORE. Furthermore, there are no significant performance drops between the validation and TC set. However, TNC appears to be much easier for both models presumably because it lacks the combined pair. Importantly, out of all the evaluation metrics used here, BLEU has the weakest (Elliott and Keller, 2014), and METEOR and CIDEr have the strongest correlations with human judgements (Yi et al., 2020). In terms of capturing semantics, SCST yields a more robust model than XE optimization as indicated by the 0.6 and 1 decline in BERTSCORE units respectively.

Considering that we kept all the conditions the same, differences must be due to the poor ability of the model to combine concepts without having observed the combinations during training, i.e. lack of systematicity. To confirm this, we inspect the Recall@K of the pairs after testing on TC. Because the combination of the pairs occur approximately

	$\mathcal{M}^2$					$\mathcal{M}^2_{SCST}$				
	R@1	R@2	R@3	R@4	R@5	R@1	R@2	R@3	R@4	R@5
black cat	1.12	2.01	3.13	3.58	4.03	1.79	1.79	2.46	3.13	4.03
big bird	0	0.81	0.81	0.81	0.81	0	0	0	0	0
red bus	10.39	14.29	19.48	22.94	27.71	12.12	13.42	15.15	16.88	17.32
small plane	0	0	0	0	0	0	0	0	0	0
eat man	11.67	13.75	17.5	21.25	22.08	8.33	8.75	10.42	12.08	12.08
lie woman	4.23	10.56	12.68	14.79	16.2	5.63	6.34	9.15	10.56	11.27
Average	4.57	6.9	8.93	10.56	11.8	4.65	5.05	6.2	7.11	7.45

Table 2: Recall scores for each pair of interest in the systematicity test.

in 1.57 of the 5 ground truth captions, it is not expected by the model to generate the combination in a single caption (Nikolaus et al., 2019). We therefore use the top 5 most likely captions generated using beam search. Table 2 illustrates recall scores for all pairs. Both models rarely perform any systematic generalizations. On average, only 4.57% (4.65%) of the time the model under XE (SCST) includes the pair in the description.

Surprisingly, SCST fine-tuning actually hinders the systematic performance of the model. While both models perform similarly when taking into account the single most likely caption, XE optimization yields significant gains by taking into account additional generations. Intuitively, SCST should facilitate exploration of the caption space. Because the reward value is a function of word overlap, for images where the majority of the reference captions do not contain the examined pair, the model will be penalized when making any systematic generalizations. This is further exacerbated by the lack of diversity in each active hypothesis during beam search decoding (Li et al., 2016; Vijayakumar et al., 2016). If most of the active hypotheses have significant overlaps with minor variations, then there is little hope for the model to make systematic generalizations in any of the K most likely generations. An alternative approach would be to modify the reward function to not penalize plausible descriptions that deviate from the ground truth. For instance, a model should not be penalized if it describes properties of objects in an image even if these properties are not mentioned in the ground truth. We leave this direction for future work.

4.2 Productivity

With productivity, we explore to what degree a captioning model can extend its predictions beyond the length distribution it has observed during training. We expose the models in two different test conditions. First, we tokenize each caption using spaCy (Honnibal et al., 2020) and compute the histogram of the average caption length for each image. The distribution of the average caption length is shown in Figure 2. For each condition, we use the same number (5000) of images for validation and testing as in Karpathy and Fei-Fei (2015). From the histogram, we assign the 5000 images with the highest caption length to the first condition, denoted Test Rich (TR). From the remaining examples, we randomly select 5000 images and assign them to the second condition - Test Base (TB), and 5000 images for validation. Lastly, the remaining $82,783$ images are used for training.

		B1	B4	M	R	C	BS
V	$\mathcal{M}^2$	76.22	36.17	28.35	57.11	115.90	44.33
	$\mathcal{M}^2_{SCST}$	81.31	39.39	29.26	59.26	130.03	45.56
TB	$\mathcal{M}^2$	76.22	35.91	28.17	56.92	116.11	44.30
	$\mathcal{M}^2_{SCST}$	80.98	39.38	29.22	59.49	130.42	45.54
TR	$\mathcal{M}^2$	75.72	35.97	24.99	53.31	85.01	40.28
	$\mathcal{M}^2_{SCST}$	80.79	39.53	26.31	55.58	95.73	41.47

Table 3: Results on productivity split in validation (V), Test Base (TB), and Test Rich (TR).

This procedure yields two independent tests, where the base test follows the same distribution of caption lengths as the train set. The rich test contains images with significantly greater length. Table 9 illustrates the average POS tags and the length of each caption per image. On average captions of images from TR have approximately 14.47% more adjectives, 31.75% more nouns, and 29.70% verbs than the train, validation, and TB.

We report the performance on the productivity test in Table 3. Overall, it appears that images containing longer captions are difficult for both models as showcased by standard captioning and semantic metrics. The performance on the TB is comparable with the validation set, however, both models perform considerably worse on the TR set. The model after XE optimization reports a drop of 31.1

CIDEr units when evaluated on longer captions. The same trend can be observed for SCST where the performance gap between TB and TR is even greater than solely training using XE. In terms of semantics, we observe a drop of approximately 4 BERTSCORE units across both methods.

	$\mathcal{M}^2$		$\mathcal{M}^2_{SCST}$	
	TB	TR	TB	TR
ADJ	0.56	0.48	0.43	0.37
ADP	1.73	1.77	1.68	1.82
ADV	0.09	0.11	0.05	0.07
CCONJ	0.15	0.19	0.20	0.24
DET	2.3	2.40	2.41	2.56
NOUN	3.42	3.49	3.45	3.62
PRON	0.03	0.03	0.04	0.04
VERB	1	1.01	0.92	0.91
LENGTH	9.36	9.58	9.34	9.74

Table 4: Average POS tags and length of generated captions in Test Base (TB) and Test Rich (TR).

Further insights can also be obtained from the average POS tags and caption length illustrated in Table 4. We observe that the model from XE optimization generates more adjectives and verbs as opposed to the model using SCST. However, the latter is generating substantially more nouns especially to describe images from the TR set. This observation also supports the findings on systematicity. If a model is generating more adjectives and verbs then it is capable of making more (adjective, noun) and (noun, verb) compositions. It is likely that the model receives greater reward by describing additional objects in the image rather than their attributes or their relations (eg 'a blue bird sitting on a bench' vs 'a bird next to two people'). As a result, the generated captions contain additional DET and ADP tags present in the caption which is also supported by the presented findings.

4.2.1 Visual Density

		B1	B4	M	R	C	BS
V	$\mathcal{M}^2$	75.89	36.38	27.83	56.65	115.0	44.06
	$\mathcal{M}^2_{SCST}$	80.74	39.43	29.08	59.01	129.11	45.38
TLD	$\mathcal{M}^2$	75.98	36.12	27.81	56.57	114.86	43.97
	$\mathcal{M}^2_{SCST}$	80.97	39.88	29.02	59.24	130.21	45.31
THD	$\mathcal{M}^2$	77.23	37.96	27.07	56.56	90.76	41.60
	$\mathcal{M}^2_{SCST}$	81.57	40.36	28.52	58.8	103.52	43.26

Table 5: Results on productivity (visual density) split in validation (V), Test Low Density (TLD), and Test High Density (THD).

The caption length may correlate with the visual information from the image and thus it may contain lots of words because the image has rich content. While this would be a nice property of captioning models, it would mean that the models do not necessarily exhibit productive behavior but simply are capable of describing additional concepts in the image. However, there is no linear dependency between the number of concepts in an image and with the length of its description (Figure 4). Consequently, a model may actually behave differently in terms of productivity if it is exposed to images with less number of objects during training.

Motivated by this observation, we repeat the productivity experiments but this time we are interested in exposing the model to images with low visual density and evaluating on images with high density. We split the dataset in a way that the test images contain significantly more numbers of concepts. Similarly, we have two test conditions: Test Low Density (TLD) and High Density (THD).

Table 5 illustrates the results in the productivity split based on image density. The word overlap evaluation metrics showcase that the models exhibit the same behavior with the previous experiments. We also observe performance drop in terms of semantic equivalence using BERTSCORE. However, it is worth mentioning that the degradation in capturing semantics is less significant than the productivity experiments using caption length. Previously, XE and CIDEr optimization recorded a difference of 4 BERTSCORE units between TLD and THD, whereas BERTSCORE declined by 2.37 and 2.05 units respectively.

4.3 Substitutivity

		B1	B4	M	R	C	BS
$\mathcal{O}$ vs $\mathcal{GT}$	$\mathcal{M}^2$	77.12	37.5	30.12	58.64	110.54	45.18
	$\mathcal{M}^2_{SCST}$	81.74	40.63	31.00	60.42	122.14	46.24
$\mathcal{S}$ vs $\mathcal{GT}$	$\mathcal{M}^2$	72.78	29.35	26.6	52.76	95.42	43.87
	$\mathcal{M}^2_{SCST}$	76.13	33.19	28.26	54.74	109.62	45.54
$\mathcal{O}$ vs $\mathcal{S}$	$\mathcal{M}^2$	62.57	35.18	32.98	60.8	304.27	65.79
	$\mathcal{M}^2_{SCST}$	77.0	55.88	44.89	75.16	466.76	77.71
$\mathcal{O}_t$ vs $\mathcal{S}_t$	$\mathcal{M}^2$	66.27	46.16	34.9	64.55	446.13	69.30
	$\mathcal{M}^2_{SCST}$	85.21	74.37	50.93	82.76	707.9	83.51

Table 6: Results on substitutivity test. ($\mathcal{GT}$) ground truth captions, ($\mathcal{O}$) original caption without substitution, ($\mathcal{S}$) caption after substitution, ($\mathcal{O}_t$) sub-caption after the synonym word, ($\mathcal{S}_t$) sub-caption after substituting the synonym word.

The objective of the final test is to evaluate the robustness of a model against synonym substitutions. In order to create a substitutivity test, we manually create two sets of words S_1, S_2. For every word

$w \in S_1$, there is another word $s \in S_2$ such that w can always be replaced by s without altering the meaning of the caption. We initially considered the 80 COCO object categories and used the mapping between objects and fine-grained classes defined by Lu et al. (2018). We excluded object categories with no synonyms ('cup') and categories containing more than one word ('baseball glove'). Next, we manually inspected ground truth captions to ensure that pairs of object categories and their synonyms are interchangeable. With this process we further divided the 'person' category into 'man', 'woman', 'boy', and 'girl' with 'person' and 'child' as synonyms. Finally, we discarded words with multiple meanings.

The pairs of object categories and synonyms used to test substitutivity are illustrated in Table 10. In order to ensure that the model is exposed to both object categories and fine-grained classes, we selected those that appeared adequate times during training. We trained a model on the train set of the Karpathy split and selected pairs of categories and fine-grained classes, where each word appears at least 200 times in the ground truth training examples. Note that in substitutivity we are not exclusively interested in images where the pair of words is jointly observed in its captions. Finally, we selected images from the test set in the Karpathy split where the generated captions of the trained model contained the fine-grained class. To verify that a substitution is performed adequate times during inference, we used pairs where the fine-grained class appeared at least 10 times in the generated captions. The distribution of the number of images with captions containing either a selected object category or fine-grained class for the train and test set are illustrated in Figure 5.

We inspect how the model behaves under replacement of a word with its synonym. During inference we apply beam search. For each active hypothesis, if the current most likely word belongs in S_1, we substitute the word with its synonym from S_2. To ensure that the substitution is preserved after each decoding stage, we set the probability of the synonym word to 1. We compute standard metrics using the original caption ($\mathcal{O}$), the caption after substitution ($\mathcal{S}$), the sub-caption after the synonym word ($\mathcal{O}_t$), and the sub-caption after replacing the synonym word ($\mathcal{S}_t$). In this setting, high values regarding overlap metrics such as BLEU and ROUGE indicate robustness of a model while semantic equivalence BERTSCORE also provides valuable insights.

The substitutivity results are illustrated in Table 6. In the first two rows we compare the ground truth caption with the originally generated caption and the caption after substitution. For both model variations we observe considerable performance drops after replacing a word with its synonym. This is expected as we intervened during decoding and replaced the original word with its synonym that had lower probability. In this case, the main concern is not whether the model generates plausible captions but whether its prediction matches the prediction before the substitution. The last two rows of the table compare the generated caption and the caption after substitution. Overall, both models performed exceptionally well with SCST providing consistent gains across all metrics. The sub-caption after substituting the synonym word appears to match with the sub-caption after the synonym word both in terms of n-gram metrics as well as semantics. This claim also holds for the captions as a whole; the high scores indicate that the models may be meaning-invariant with regards substitutions from fine-grained classes to more generic ones.

Our findings suggest that the model is robust against these substitutions. However, it may be straightforward for the model to substitute a fine-grained object description with another that has a broader concept. A more challenging scenario would involve the same experiment but replacing generic descriptions with more fine-grained categories. For instance, replacing 'person' in 'A person driving truck' with 'firefighter'.

5 Qualitative Analysis

For each proposed test we randomly sampled 100 examples from the derived splits and inspected the generated captions. In this section, we report the main findings based on that pool. Additional material is provided in the Appendix C.

Systematicity We observed that the models from both training procedures are reluctant to make systematic generalizations. In the case of adjective and noun pairs the models consistently avoided using adjectives or used adjectives that describe a different property of the object. In these cases the models do not actually learn to combine pairs but instead learn co-occurrence statistics in the data (e.g., 'a double decker bus' and 'a red bus'). With regards to pairs of nouns and verbs the models tended to replace the verb with a generic phrase. We also found an adequate number of examples where the generated caption did not contain any verb at all.

Productivity Overall, both models favored short captions. We observed cases where the ground truth captions provided fine-grained explanations, yet the generated caption contained only a handful of these descriptions. However, this does not entail that the generated caption is incorrect or of poor quality. Both models generally performed reasonably well, without hallucinating objects in the scenes or assigning incorrect properties to described objects. We also frequently observed cases where at least one of the reference captions constitutes an outlier in terms of caption length. In these cases the annotator provided a thorough description of the image. To maximize its performance, the model prioritizes matching the generated with the remaining reference captions whose lengths cluster around similar values.

Substitutivity In the cases of mismatch between the originally generated caption and its modification, the majority of the examples differed exclusively in the part of the caption after the substitution. The modified caption either contained the same objects and their attributes with a simple reordering or the objects were described with more detail including additional properties or relations. We observed a few examples where the caption was completely restructured and identified two cases of such behavior. On the one hand, the original caption contained multiple occurrences of fine-grained objects (e.g., 'a man and a woman riding on a motorcycle' & 'a person riding a motorcycle with a person on the back'). On the other hand, the caption was altered to include additional properties of the substituted word (e.g., 'a living room with a television and a fireplace' & 'a flat screen tv in a living room with a fireplace'). These cases could be due to the decoding policy as substituting the original word with its synonym in an active hypothesis results in a sequence with lower marginal probability. The active hypothesis is then discarded as it does not fit in the beam width. This is a common problem in decoding, where high probability words are concealed behind low probability words.

6 Conclusion

We presented a series of tests for compositionality in image captioning. This work contributes towards what it means for a captioning model to be 'compositional', and what properties we would like them to have. We performed data and model ablations to identify limitations of state of the art models across three dimensions of compositionality.

Our findings in the systematicity align with the findings from previous works. We find that transformer-based captioning models rarely make systematic generalizations. However, as shown by the experiments in productivity, this is also partially due to the model not producing adjectives and nouns. We demonstrated that the well-established CIDEr fine-tuning coupled with beam search decoding actually exacerbates the already poor performance on systematicity.

In productivity, we found out that models struggle to extend their predictions to match the length of the ground truth captions. Both models trained using XE and SCST generated less number of adjectives, nouns, and verbs compared to the ground truth captions. On average we observed that models after XE optimization provide captions with more adjectives and verbs, while models incorporating SCST generate descriptions with more nouns. We further included a set of experiments concerning the visual density of the image with similar results.

The substitutivity experiment showcased that it is easy for the models to substitute a fine-grained with abstract descriptions of concepts. In most cases, the part of the caption following the synonym word was identical to the part of the caption after the substitution with its synonym. A natural extension to the substitutivity experiment would include the performance after substituting more abstract with fine-grained descriptions.

With our framework we provided insights regarding the evaluation and training of image captioning models. Word overlap metrics favor models that generate sequences closer to the target rather than more 'grounded' models that focus on actual properties of the objects in the image. This calls for a training regime that mitigates this issue by introducing multimodal metrics that take both text and vision into account (e.g., *CLIPScore* (Hessel et al., 2021)). Additionally, different training strategies should be adopted to allow the model to explore the search space, and learn to generate sequences that go beyond the average sequence length.

7 Ethical statement

The presented paper introduces a framework to evaluate compositionality of image captioning models from multiple perspectives. The dataset and the model under evaluation are publicly available for academic purposes and not intended for downstream deployment.

Despite recent advances, our findings challenge the systematicity and productivity of current models. This suggests that the generalization capacity and robustness remain a barrier to overcome, before exposing the outputs of these models to end users. As a result, we believe that comprehensive evaluations can help expose biases in the model and minimize the impact in real-world deployment of language technologies.

Acknowledgements

This research was supported by Alana AI. The views, and/or opinions presented are those of the authors and not of the funding agency.

References

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*.

Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. 2020. Learning to recombine and resample data for compositional generalization. In *International Conference on Learning Representations*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. 2016. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2018. Systematic generalization: What is required and can it be learned? In *International Conference on Learning Representations*.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Emanuele Bugliarello and Desmond Elliott. 2021. The role of syntactic planning in compositional image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 593–607.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Henry Conklin, Bailin Wang, Kenny Smith, and Ivan Titov. 2021. Meta-learning to compositionally generalize. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3322–3335, Online. Association for Computational Linguistics.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics.

Jerry A Fodor and Ernest Lepore. 2002. *The compositionality papers*. Oxford University Press.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Gottlob Frege. 1950. *The foundations of arithmetic: A logico-mathematical enquiry into the concept of number*. Northwestern University Press.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: how do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Yichen Jiang and Mohit Bansal. 2021. Inducing transformer's compositional generalization ability via auxiliary sequence prediction tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6253–6265.

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.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2873–2882. PMLR.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Association for Computational Linguistics Human Language Technology Conference*, pages 236–244.

Mathijs Mul and Willem Zuidema. 2019. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization. *arXiv preprint arXiv:1906.00180*.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational*

Natural Language Learning (CoNLL), pages 87–98. Association for Computational Linguistics.

Peter Pagin. 2003. Communication and strong compositionality. *Journal of Philosophical Logic*, 32(3):287–322.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matthew Purver, Mehrnoosh Sadrzadeh, Ruth Kempson, Gijs Wijnholds, and Julian Hough. 2021. Incremental composition in distributional semantics. *Journal of Logic, Language and Information*, 30(2):379–406.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Paweł Krzysztof Nowak, Tal Linzen, Fei Sha, and Kristina Toutanova. 2021. Improving compositional generalization with latent structure and data augmentation. *arXiv preprint arXiv:2112.07610*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. 2020. A benchmark for systematic generalization in grounded language understanding. *Advances in neural information processing systems*, 33:19861–19872.

Mehrnoosh Sadrzadeh and Edward Grefenstette. 2011. A compositional distributional semantics, two concrete constructions, and some experimental evaluations. In *International Symposium on Quantum Interaction*, pages 35–47. Springer.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216.

Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon. 2020. CompGuessWhat?!: A multi-task evaluation framework for grounded language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7625–7641, Online. Association for Computational Linguistics.

Zoltan Szabó. 2012. The case for compositionality. *The Oxford handbook of compositionality*, 64:80.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Sara Veldhoen, Dieuwke Hupkes, and Willem H Zuidema. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@ NIPS*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. 2021. Separating skills and concepts for novel visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5632–5641.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, Online. Association for Computational Linguistics.

Pengcheng Yin, Hao Fang, Graham Neubig, Adam Pauls, Emmanouil Antonios Platanios, Yu Su, Sam Thomson, and Jacob Andreas. 2021. Compositional generalization for neural semantic parsing via span-level supervised attention. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Online. Association for Computational Linguistics.

A Dataset splits

We created custom splits of the MSCOCO dataset (Lin et al., 2014), a collection of images described in English.

A.1 Systematicity

black cat
big bird
red bus
small plane
eat man
lie woman

Table 7: Pairs of concepts used to test systematicity.

Recall scores

	Color	Size	Verb
BUTD Nikolaus et al.	15.95	0.32	10.55
$\mathcal{M}^2$	15.78	0.41	19.14
$\mathcal{M}^2_{SCST}$	8.66	0	11.7

Table 8: Recall@5 for each grouped category of concepts of interest.

Additional insights can be obtained by observing the individual recall scores for each concept of interest. Both models cannot make systematic generalizations in terms of adjectives describing size. This aligns with the view of (Nikolaus et al., 2019) who also showcased that the actual bounding box of the referred noun does not correlate with its size modifiers in the description of an image. For additional comparison with the work of Nikolaus et al. (2019), we group the pairs in terms of color, size, and verb as shown in Table 8. Interestingly, our findings suggest that transformer-based architectures are more capable of systematic composition when they describe verbs. On the other hand, BUTD recorded the best generalization performance when they describe color and noun pairs. There is no reported performance of BUTD for the systematicity split using SCST.

A.2 Productivity

The distribution of the average caption length is shown Figure 2. An overview of the average POS tags and the length of each caption per image is illustrated in Table 9, where the left and right part of the table account for the Karpathy and the proposed productivity split. On average captions of images from Test Rich have approximately 14.47% more adjectives, 31.75% more nouns, and 29.70% verbs than the train, validation, and test base sets. We apply the same procedure for the visual density experiment. Figure 3 shows the distribution of the number of instances on each split.

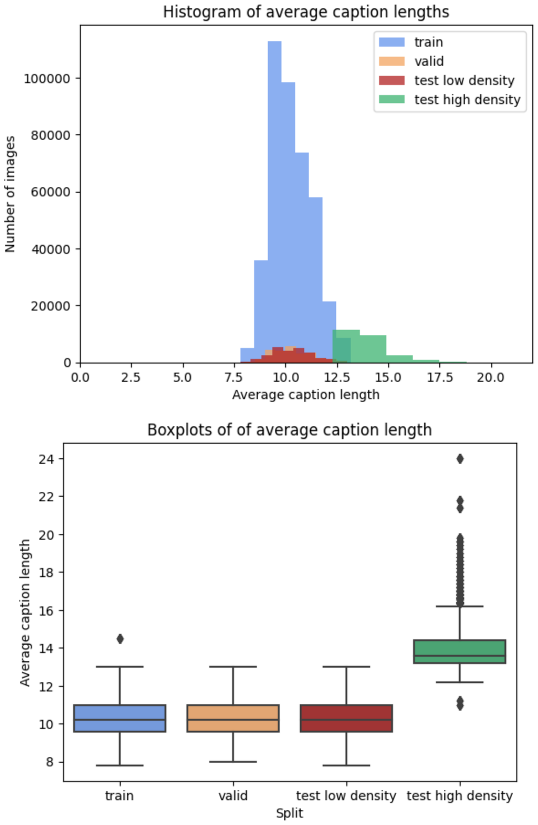


Figure 2: Histogram and boxplots of average caption length for each image in the train, validation, and test sets of the productivity split.

A.3 Substitutivity

We initially considered the 80 COCO categories and used the mapping between objects and fine-grained classes defined by Lu et al. (2018). We excluded object categories with no synonyms ('cup') and categories containing more than one word ('baseball glove'). Next, we manually inspected ground truth captions to ensure that pairs of object categories and their synonyms are interchangeable. With this process we further divided the 'person' category into 'man', 'woman', 'boy', and 'girl' with 'person' and 'child' as synonyms. Finally, we discarded the 'dog' category completely as we found that it often referred to the actual animal or 'hot dog'.

	Train	Valid	Test	Train	Valid	Test Base	Test Rich
ADJ	0.76	0.76	0.77	0.76	0.77	0.76	0.87
ADP	1.74	1.75	1.75	1.71	1.71	1.71	2.46
ADV	0.15	0.15	0.16	0.15	0.15	0.14	0.21
CCONJ	0.24	0.24	0.24	0.23	0.23	0.23	0.45
DET	2.2	2.21	2.2	2.17	2.17	2.18	2.9
NOUN	3.64	3.64	3.62	3.59	3.58	3.58	4.73
PRON	0.18	0.18	0.18	0.17	0.18	0.17	0.32
VERB	1.02	1.02	1.01	1.01	1	1.01	1.31
LENGTH	11.34	11.35	11.32	11.19	11.17	11.2	15.03

Table 9: Comparison of average POS tags and caption lengths in each image between train, validation, and test sets in the Karpathy (Karpathy and Fei-Fei, 2015) and the productivity split.

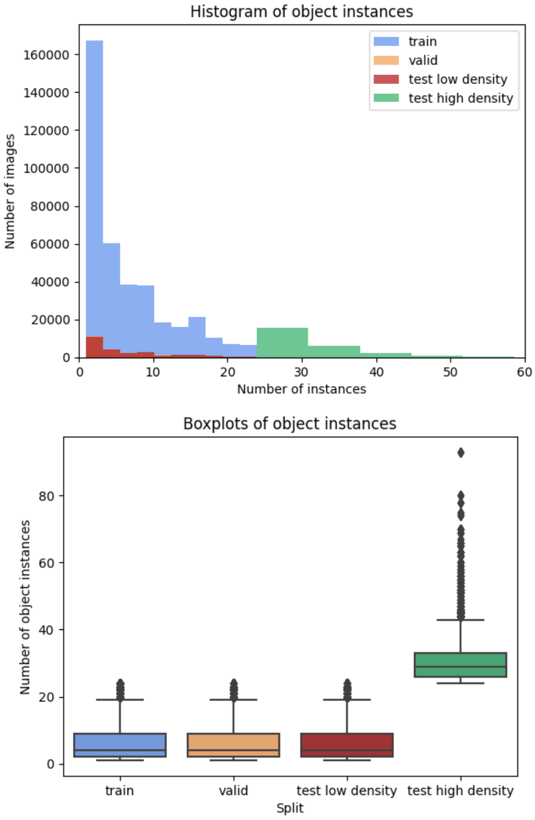


Figure 3: Histogram and boxplots of number of instances for each image in the train, validation, and test sets of the productivity split concerning visual density.

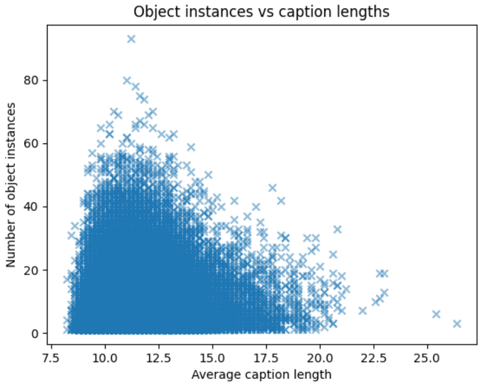


Figure 4: Illustration of the number of instances and the caption length of images in MSCOCO (Lin et al., 2014).

Object category	Fine-grained class
`person`	`man`, `woman`
`child`	`boy`, `girl`
`bicycle`	`bike`
`airplane`	`plane`, `jet`, `jetliner`
`cow`	`cattle`
`tv`	`television`
`refrigerator`	`freezer`
`laptop`	`computer`

Table 10: Selected pairs of object categories and fine-grained classes used in substitutivity split. During inference, we replace the generated fine-grained word with its synonym.

B Model details

Our implementation is based on the publicly available PyTorch codebase of $\mathcal{M}^2$-transformer (`https://github.com/aimagelab/meshed-memory-transformer`). Following Cornia et al. (2020) we use 3 encoding and decoding layers, 8 attention heads, and 40 memory vectors.

We also noticed during SCST, that the model occasionally produced incomplete captions (e.g., 'a man is riding a horse in a'). Our interpretation here is that the model is reluctant to produce that noun and the learnt policy indicates that it is better to generate an incomplete caption and receive the adjusted reward rather than make a 'risky' prediction. The paper introducing SCST (Rennie et al., 2017) states in the supplementary materials (sec-

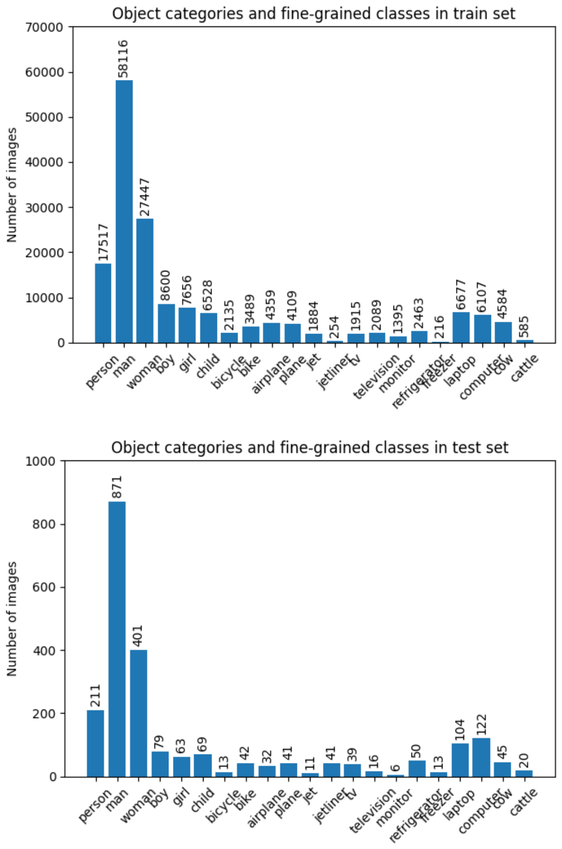


Figure 5: Distribution of object categories and fine-grained classes in train (left) and test (right) substitutivity split.

tion E): "*One detail that was crucial to optimizing CIDEr to produce better models was to include the EOS tag as a word*." If the EOS word is omitted, trivial sentence fragments such as 'with a' and 'and a' receive significant reward values, as opposed to their full sentence counterparts. However, including the EOS tag lowers the reward allocated to the incomplete captions. We apply the same procedure by appending EOS token to both candidate and reference captions.

C Qualitative analysis

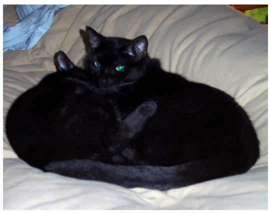
We examined qualitatively the behavior of the models under each compositional test by randomly sampling 100 examples. For systematicity (Figure 6), we compared the occurrences of systematic generalization in Test Comb. Similarly, for productivity (Figure 7) we studied the captions over images belonging to Test Rich with a focus on the part-of-speech tags produced during generation as opposed to the reference captions. Finally, for substitutivity (Figure 8) we examined the originally generated and the modified caption and emphasized on their similarity.

$\mathcal{GT}$: a cute cat sticking its head in a box of pizza,
a white and black cat with its head inside a box smelling the food,
a cat pokes its head into a box and smells the food inside it,
a cat with its head in a box of pizza,
a cat trying to sneak a bite of pizza
$\mathcal{M}^2$: a white and **black cat** eating a piece of pizza
$\mathcal{M}^2_{SCST}$: a cat is eating a pizza in a box

$\mathcal{GT}$: two red buses headed to the same place are right next to each other on the road,
buses lined up on the street in traffic,
there are many red busses coming down the street together,
the buses are lined up waiting for passengers,
a couple of buses drive next to each other
$\mathcal{M}^2$: a couple of **red buses** driving down a street
$\mathcal{M}^2_{SCST}$: two **red buses** driving down a city street

$\mathcal{GT}$: a bright blue and white amx jet is in the clear sky,
a blue airplane is flying during a clear day,
an airplane flying in a blue sky,
a small two toned blue airplane flying,
a small plane is seen flying on a clear day
$\mathcal{M}^2$: a blue and white airplane flying in the sky
$\mathcal{M}^2_{SCST}$: an airplane is flying in the blue sky

$\mathcal{GT}$: a man holding a slice of pizza while wearing glasses,
there is a man eating a sandwich with lots of cheese on it,
a man in red is eating some food,
a full view of an individual in the image,
a man looks at the camera while holding a hot dog
$\mathcal{M}^2$: a man in a red shirt holding two hot dogs
$\mathcal{M}^2_{SCST}$: a man in a red shirt holding a hot dog



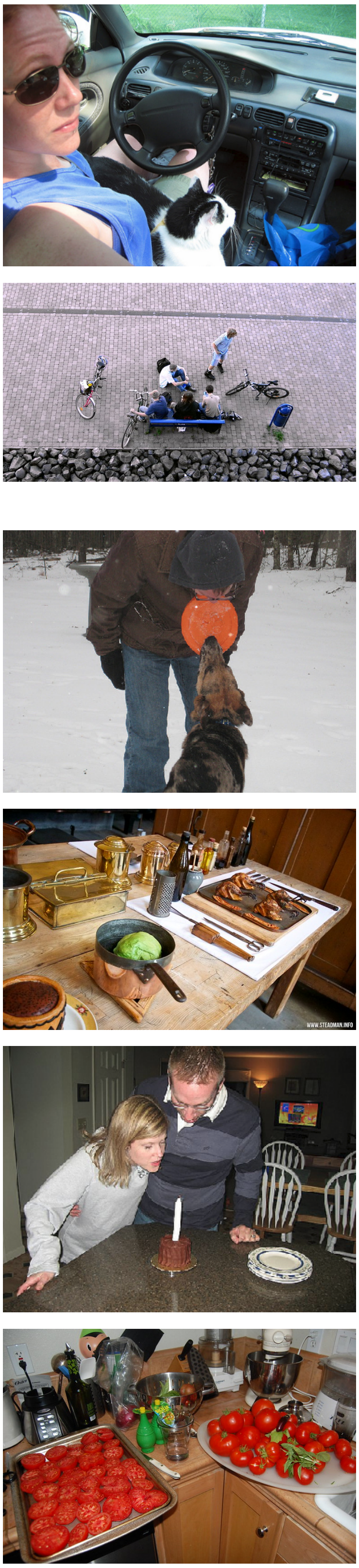
$\mathcal{GT}$: a couple of black cats laying on top of a bed,
two black cats cuddle together on a blanket,
two black cats sleeping together on a bed,
two black cats cuddled together on a bed,
a couple of cats relaxing with each other on the bed
a woman sitting in the drivers seat of a car with a cat in her lap
$\mathcal{M}^2$: a cat laying on a blanket on a bed
$\mathcal{M}^2_{SCST}$: a **black cat** laying on a bed

$\mathcal{GT}$: a red bus on street next to buildings,
a public transit bus on a city street,
a large red bus on a city street,
a red bus crossing a street next to tall buildings,
a red bus is parked along the side of a street
$\mathcal{M}^2$: a double decker bus driving down a city street
$\mathcal{M}^2_{SCST}$: a double decker bus driving down a city street

$\mathcal{GT}$: an airplane with wheels wheels barely off ground tilted slightly upward from the pavement to the blue sky,
a small plane is taking off from a sandy beach,
a white airplane is driving down the runway,
small plane inches above flat surface near water,
a small plane on the sand near a beach
$\mathcal{M}^2$: an airplane is on the runway on a sunny day
$\mathcal{M}^2_{SCST}$: an airplane is taking off from an airport runway

$\mathcal{GT}$: three guys sitting down eating sandwiches and smiling,
three men eating sandwiches at a corner table,
two young men one old enjoying a meal at a restaurant,
three men all eating sub sandwiches at a restaurant,
three men are sitting in a restaurant eating sandwiches
$\mathcal{M}^2$: three **men** sitting at a table **eating** food
$\mathcal{M}^2_{SCST}$: three **men** sitting at a table **eating** a sandwich

Figure 6: Examples of generated captions for different concept pairs from the Test Comb. Bold phrases indicate successful systematic generalization.



$\mathcal{GT}$: a woman driving a car while holding a cat on her lap
a woman driving her car with a cat riding in her lap,
a lady driving her car with a black and white cat in her lap,
a woman sitting in a car with a black and white cat,
a woman sitting in the drivers seat of a car with a cat in her lap
$\mathcal{M}^2$: a person in a car with a cat
$\mathcal{M}^2_{SCST}$: a woman in a car with a black and white cat

$\mathcal{GT}$: a bunch of people sitting on and standing around a bench with bikes,
some people sit on a bench near bicycles,
a group of people sit on and near a park bench,
several people sit on a blue bench with their bikes around them,
a bench seats a few people as bikes are parked nearby and one man sits on a brick walkway as another boy in blue stands near them
$\mathcal{M}^2$: a couple of people sitting on top of a bench
$\mathcal{M}^2_{SCST}$: a bike with a bench and people in it

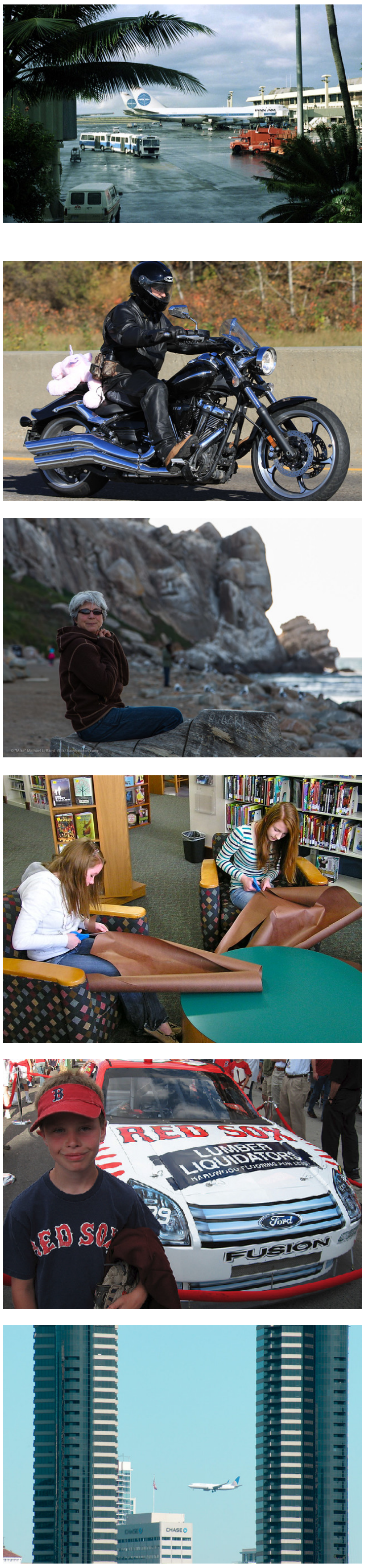
$\mathcal{GT}$: a man holding an orange frisbee in his mouth with a dog,
a dog and its owner battling over a frisbee,
a person with a frisbee in his mouth bending over to his dog who has the other end of the frisbee in its mouth,
a man in the snow holding a disc in his mouth as a dog bites it also,
a man and dog use their teeth to fight for the same frisbee
$\mathcal{M}^2$: a man holding an orange dog in the snow
$\mathcal{M}^2_{SCST}$: a man holding an orange frisbee with a dog

$\mathcal{GT}$: a wooden kitchen table topped with baked goods and pie,
a tray with some food a pot and some bottles,
there is a pan with lettuce in it near a tray of meat,
a tray of food and a boiler with a vegetable sit on a kitchen counter,
a counter with a pot with a vegetable in it as well as chicken breasts on the side
$\mathcal{M}^2$: a wooden cutting board topped with lots of food
$\mathcal{M}^2_{SCST}$: a wooden table with a pan of food and a knife

$\mathcal{GT}$: father and daughter leaning over small cake with large candle on it,
a man and a woman blowing out a candle in a cake,
a guy and girl celebrating an occasion with a cake with chocolate frosting and 1 candle,
a man and woman stand before a small cake with a single candle in it,
a couple blowing out an enormous candle on a small chocolate
$\mathcal{M}^2$: a man blowing out candles on a birthday cake
$\mathcal{M}^2_{SCST}$: a man and a woman blowing out candles on a birthday cake

$\mathcal{GT}$: a kitchen counter top with a tray of sliced tomatoes and a plate of whole tomatoes,
there is a large plate of tomatoes and a pan of sliced tomatoes,
a cookie sheet with red sliced tomatoes and a platter of whole tomatoes on a crowded kitchen counter, there 's plenty of red tomatoes on the kitchen counter,
a sloe up of sliced tomatoes on a baking pan
$\mathcal{M}^2$: a close up of a plate of food with tomatoes
$\mathcal{M}^2_{SCST}$: a kitchen counter with a bunch of tomatoes and other vegetables

Figure 7: Productivity: examples of generated captions from images in the Test Rich.



$\mathcal{GT}$: an airport with large jetliners and a bus traveling on a tarmac,
an airplane and busses are lined up at the airport,
a group of buses driving around at the airport,
airplanes sit at the gate as transportation vehicles move about,
a busy runway with buses and luggage carts driving around
$\mathcal{M}^2$: a large jetliner sitting on top of an airport tarmac
$\mathcal{M}^2$ (S): a large airplane that is on a runway
$\mathcal{M}^2_{SCST}$: a plane is parked at an airport terminal
$\mathcal{M}^2_{SCST}$ (S): a airplane parked at an airport with cars and planes

$\mathcal{GT}$: a person sits on top of a motorcycle with a stuffed toy,
a person riding a motorcycle with a stuffed animal on the back,
a person on a motorcycle with a stuffed animal on back,
a motorcyclist riding with a stuffed animal attached to the back,
a person in full leather riding a motorcycle with a stuff animal on the back
$\mathcal{M}^2$: a man riding on the back of a motorcycle
$\mathcal{M}^2$ (S): a person riding a motorcycle on a street
$\mathcal{M}^2_{SCST}$: a man riding a motorcycle on a road
$\mathcal{M}^2_{SCST}$ (S): a person riding a motorcycle on a road

$\mathcal{GT}$: an older woman sits in a sweater at the beach,
a person wearing sun glasses and blue jeans sitting on a rock by the ocean,
a woman is sitting on the beach,
a lady near some rocks during the daytime looking at the camera,
an older woman sitting on a drift log at a beach
$\mathcal{M}^2$: a woman sitting on a log near the water
$\mathcal{M}^2$ (S): an older person sitting on a log in front of a mountain
$\mathcal{M}^2_{SCST}$: a woman sitting on a log by the water
$\mathcal{M}^2_{SCST}$ (S): an older person sitting on a log near the water

$\mathcal{GT}$: two girls in a library seated at a table cutting large brown paper,
girls sitting in a library cutting brown paper,
two girls working on a project in the library,
a couple of girls cutting paper with some scissors,
two teenaged girls sitting in armchairs at a public library and cutting sheets of craft paper with scissors $\mathcal{M}^2$: two girls sitting on chairs in a library
$\mathcal{M}^2$ (S): two young children sitting together in a library
$\mathcal{M}^2_{SCST}$: two girls sitting in chairs in a library
$\mathcal{M}^2_{SCST}$ (S): two children sitting in chairs in a library

$\mathcal{GT}$: a young man standing next to a race car with the red sox logo on it 's hood,
a young boy standing in front of a sponsored car,
a man standing near a red sox nascar,
a young boy standing by a red sox car wearing red sox shirt and visor,
a young man standing next to a racecar on a display lot
$\mathcal{M}^2$: a young boy wearing a red hat standing in front of a car
$\mathcal{M}^2$ (S): a young child standing in front of a car
$\mathcal{M}^2_{SCST}$: a young boy is standing next to a car
$\mathcal{M}^2_{SCST}$ (S): a young child standing next to a police car

$\mathcal{GT}$: a low flying commercial plane passing tall buildings,
an airplane is flying in the sky beyond some skyscrapers,
a jetliner flying low as viewed between two skyscrapers,
an airplane is seen in the air between two buildings,
an airplane flying pass building and a bank building
$\mathcal{M}^2$: a large jetliner flying over a tall building
$\mathcal{M}^2$ (S): a large airplane flying over a city skyline
$\mathcal{M}^2_{SCST}$: a large jetliner flying over a tall building
$\mathcal{M}^2_{SCST}$ (S): a large airplane flying over a city skyline

Figure 8: Substitutivity: examples of generated captions from images in the substitutivity test.