Compositional Generalization in Image Captioning

Mitja Nikolaus  
University of Tübingen*  
mitja.nikolaus@posteo.de

Mostafa Abdou  
University of Copenhagen  
abdou@di.ku.dk

Matthew Lamm  
Stanford University  
mlamm@cs.stanford.edu

Rahul Aralikatte  
University of Copenhagen  
rahul@di.ku.dk

Desmond Elliott  
University of Copenhagen  
de@di.ku.dk

Abstract

Image captioning models are usually evaluated on their ability to describe a held-out set of images, not on their ability to generalize to unseen concepts. We study the problem of compositional generalization, which measures how well a model composes unseen combinations of concepts when describing images. State-of-the-art image captioning models show poor generalization performance on this task. We propose a multi-task model to address the poor performance, that combines caption generation and image–sentence ranking, and uses a decoding mechanism that re-ranks the captions according to their similarity to the image. This model is substantially better at generalizing to unseen combinations of concepts compared to state-of-the-art captioning models.

1 Introduction

When describing scenes, humans are able to almost arbitrarily combine concepts, producing novel combinations that they have not previously observed (Matthei, 1982; Piantadosi and Aslin, 2016). Imagine encountering a purple-colored dog in your town, for instance. Given that you understand the concepts PURPLE and DOG, you are able to compose them together to describe the dog in front of you, despite never having seen one before.

Image captioning models attempt to automatically describe scenes in natural language (Bernardi et al., 2016). Most recent approaches generate captions using a recurrent neural network, where the image is represented by features extracted from a Convolutional Neural Network (CNN). Although state-of-the-art models show good performance on challenge datasets, as measured by text-similarity metrics, their performance as measured by human judges is low when compared to human-written captions (Vinyals et al., 2017, Section 5.3.2).

It is widely believed that systematic compositionality is a key property of human language that is essential for making generalizations from limited data (Montague, 1974; Partee, 1984; Lake et al., 2017). In this work, we investigate to what extent image captioning models are capable of compositional language understanding. We explore whether these models can compositionally generalize to unseen adjective–noun and noun–verb composition pairs, in which the constituents of the pair are observed during training but the combination is not, thus introducing a paradigmatic gap in the training data, as illustrated in Figure 1. We define new training and evaluation splits of the COCO dataset (Chen et al., 2015) by holding out the data associated with the compositional pairs from the training set. These splits are used to evaluate how well models generalize to describing images that depict the held out pairings.

Figure 1: We evaluate whether image captioning models are able to compositionally generalize to unseen combinations of adjectives, nouns, and verbs by forcing paradigmatic gaps in the training data.

We find that state-of-the-art captioning models, such as Show, Attend and Tell (Xu et al., 2015), and Bottom-Up and Top-Down Attention (Ander-
son et al., 2018), have poor compositional generalization performance. We also observe that the inability to generalize of these models is primarily due to the language generation component, which relies too heavily on the distributional characteristics of the dataset and assigns low probabilities to unseen combinations of concepts in the evaluation data. This supports the findings from concurrent work (Holtzman et al., 2019) which studies the challenges in decoding from language models trained with a maximum likelihood objective.

To address the generalization problem, we propose a multi-task model that jointly learns image captioning and image–sentence ranking. For caption generation, our model benefits from an additional step, where the set of captions generated by the model can be re-ranked using the jointly-trained image–sentence ranking component. We find that the ranking component is less affected by the likelihood of n-gram sequences in the training data, and that it is able to assign a higher ranking to more informative captions which contain unseen combinations of concepts. These findings are reflected by improved compositional generalization.

The source code is publicly available on GitHub.\(^1\)

2 Related Work

2.1 Caption Generation and Retrieval

Image Caption Generation models are usually end-to-end differentiable encoder-decoder models trained with a maximum likelihood objective. Given an image encoding that is extracted from a convolutional neural network (CNN), an RNN-based decoder generates a sequence of words that form the corresponding caption (Vinyals et al., 2015, inter-alia). This approach has been improved by applying top-down (Xu et al., 2015) and bottom-up attention mechanisms (Anderson et al., 2018). These models show increasingly good performance on benchmark datasets, e.g. COCO, and in some cases reportedly surpass human-level performance as measured by n-gram based evaluation metrics (Bernardi et al., 2016). However, recent work has revealed several caveats. Firstly, when using human judgments for evaluation, the automatically generated captions are still considered worse in most cases (Fang et al., 2015; Vinyals et al., 2017). Furthermore, when evaluating out-of-domain images or images with unseen concepts, it has been shown that the generated captions are often of poor quality (Mao et al., 2015; Vinyals et al., 2017). Attempts have been made to address the latter issue by leveraging unpaired text data or pre-trained language models (Hendricks et al., 2016; Agrawal et al., 2018).

Image–Sentence Ranking is closely related to image captioning. Here, the problem of language generation is circumvented and models are instead trained to rank a set of captions given an image, and vice-versa (Hodosh et al., 2013). A common approach is to learn a visual–semantic embedding for the captions and images, and to rank the images or captions based on similarity in the joint embedding space. State-of-the-art models extract image features from CNNs and use gated RNNs to represent captions, both of which are projected into a joint space using a linear transformation (Frome et al., 2013; Karpathy and Fei-Fei, 2015; Vendrov et al., 2016; Faghri et al., 2018).

2.2 Compositional Models of Language

Investigations of compositionality in vector space models date back to early debates in the cognitive science (Fodor and Pylyshyn, 1988; Fodor and Lepore, 2002) and connectionist literature (McClelland et al., 1986; Smolensky, 1988) regarding the ability of connectionist systems to compose simple constituents into complex structures. In the NLP literature, numerous approaches that (loosely) follow the linguistic principle of compositionality\(^2\) have been proposed (Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011). More recently, it has become standard to employ representations which are learned using neural network architectures. The extent to which these models behave compositionally is an open topic of research (Lake and Baroni, 2017; Dasgupta et al., 2018; Ettinger et al., 2018; McCoy et al., 2018) that closely relates to the focus of the present paper.

Compositional generalization in image captioning has received limited attention in the literature. In Atzmon et al. (2016), the captions in the COCO dataset are replaced by subject-relation-object triplets, circumventing the problem of language generation, and replacing it with structured

\(^{1}\)https://github.com/mitjanikolaus/compositional-image-captioning

\(^{2}\)The meanings for complex expressions are derived from the meanings of their parts via specific composition functions. (Partee, 1984)
triplet prediction. Other work explores generalization to unseen combinations of visual concepts as a classification task (Misra et al., 2017; Kato et al., 2018). Lu et al. (2018) is more closely related to our work; they evaluate captioning models on describing images with unseen noun-noun pairs.

In this paper, we study compositional generalization in image captioning with combinations of multiple classes of nouns, adjectives, and verbs. We find that state-of-the-art models fail to generalize to unseen combinations, and present a multi-task model that improves generalization by combining image captioning (Anderson et al., 2018) and image–sentence ranking (Faghri et al., 2018). In contrast to other models that use a re-ranking step, our model is trained jointly on both tasks and does not use any additional features or external resources. The ranking model is only used to optimize the global semantics of the generated captions with respect to the image.

3 Compositional Image Captioning

3.1 Problem Definition

In this section we define the compositional captioning task, which is designed to evaluate how well a model generalizes to captioning images that should be described using previously unseen combinations of concepts, when the individual concepts have been observed in the training data.

We assume a dataset of captioned images \( D \), in which \( N \) images are described by \( K \) captions: \( D := \{ (i^1, s^1_n), ..., (i^K, s^K_n) \} \). We also assume the existence of a concept pair \( \{ c_i, c_j \} \) that represents the concepts of interest in the evaluation. In order to evaluate the compositional generalization of a model for that concept pair, we first define a training set by identifying and removing instances where the captions of an image contain the pair of concepts, creating a paradigmatic gap in the original training set: \( D_{\text{train}} := \{ (i^n, s^n_k) \} \) s.t. \( \forall_{n=1}^N \exists k : c_i \in s^n_k \land c_j \in s^n_k \). Note that the concepts \( c_i \) and \( c_j \) can still be independently observed in the captions of an image of this set, but not together in the same caption. We also define validation and evaluation sets \( D_{\text{val}} \) and \( D_{\text{eval}} \) that only contain instances where at least one of the captions of an image contains the pair of concepts: \( D_{\text{val/eval}} := \{ (i^n, s^n_k) \} \) s.t. \( \forall_{n=1}^N \exists k : c_i \in s^n_k \land c_j \in s^n_k \). A model is trained on the \( D_{\text{train}} \) training set until it converges, as measured on the \( D_{\text{val}} \) validation set. The compositional generalization of the model is measured by the proportion of evaluation set captions which successfully combined a held out pair of concepts \( \{ c_i, c_j \} \) in \( D_{\text{eval}} \).

3.2 Selection of Concept Pairs

We select pairs of concepts that are likely to be represented in an image recognition model. In particular, we identify adjectives, nouns, and verbs in the English COCO captions dataset (Chen et al., 2015) that are suitable for testing compositional generalization. We define concepts as sets of synonyms for each word, to account for the variation in how the concept can be expressed in a caption. For each noun, we use the synonyms defined in Lu et al. (2018). For the verbs and adjectives, we use manually defined synonyms (see Appendix D). From these concepts, we select adjective-noun and noun-verb pairs for the evaluation. To identify concept pair candidates, we use StanfordNLP (Qi et al., 2018) to label and lemmatize the nouns, adjectives, and verbs in the captions, and to check if the adjective or verb is connected to the respective noun in the dependency parse.

Nouns: We consider the 80 COCO object categories (Lin et al., 2014) and additionally divide the “person” category into “man”, “woman” and “child”. It has been shown that models can detect and classify these categories with high confidence (He et al., 2016). We further group the nouns under consideration into animate and inanimate objects. We use the following nouns in the evaluation: woman, man, dog, cat, horse, bird, child, bus, plane, truck, table.

Adjectives: We analyze the distribution of the adjectives in the dataset (see Figure 4 in Appendix A). The captions most frequently contain descriptions of the color, size, age, texture or quantity of objects in the images. We consider the color and size adjectives in this evaluation. It has been shown that CNNs can accurately classify the color of objects (Anderson et al., 2016); and we assume that CNNs can encode the size of objects because they can predict bounding boxes, even for small
Table 1: The 24 concept pairs used to construct the training $D_{train}$ and eval $D_{eval}$ datasets.

<table>
<thead>
<tr>
<th>black cat</th>
<th>big bird</th>
<th>red bus</th>
</tr>
</thead>
<tbody>
<tr>
<td>small plane</td>
<td>eat man</td>
<td>lie woman</td>
</tr>
<tr>
<td>white truck</td>
<td>small cat</td>
<td>brown dog</td>
</tr>
<tr>
<td>big plane</td>
<td>ride woman</td>
<td>fly bird</td>
</tr>
<tr>
<td>white horse</td>
<td>big cat</td>
<td>blue bus</td>
</tr>
<tr>
<td>small table</td>
<td>hold child</td>
<td>stand bird</td>
</tr>
<tr>
<td>black bird</td>
<td>small dog</td>
<td>white boat</td>
</tr>
<tr>
<td>stand child</td>
<td>big truck</td>
<td>eat horse</td>
</tr>
</tbody>
</table>

Verbs: Sadeghi and Farhadi (2011) show that it is possible to automatically describe the interaction of objects or the activities of objects in images. We select verbs that describe simple and well-defined actions and group them into transitive and intransitive verbs. We use the following verbs in the pairs: eat, lie, ride, fly, hold, stand.

Pairs and Datasets: We define a total of 24 concept pairs for the evaluation, as shown in Table 1. The training and evaluation data is extracted from the COCO dataset, which contains $K=5$ reference captions for $N=123,287$ images. In the compositional captioning evaluation, we define the training datasets $D_{train}$ and validation datasets $D_{val}$ as subsets of the original COCO training data, and the evaluation datasets $D_{eval}$ as subsets of the COCO validation set, both given the concept pairs.

To ensure that there is enough evaluation data, we only use concept pairs for which there are more than 100 instances in the validation set. Occurrence statistics for the considered concept pairs can be found in Appendix B.

3.3 Evaluation Metric

The performance of a model is measured on the $D_{eval}$ datasets. For each concept pair evaluation set consisting of $M$ images, we dependency parse the set of $M \times K$ generated captions $\{s_1, ..., s_K\}, ..., \{s'_1, ..., s'_K\}$ to determine whether the captions contain the expected concept pair, and whether the adjective or verb is a dependent of the noun.\(^5\) We denote the set of captions for which these conditions hold true as $C$.

There is low inter-annotator agreement in the human reference captions on the usage of the concepts in the target pairs.\(^6\) Therefore, one should not expect a model to generate a single caption with the concepts in a pair. However, a model can generate a larger set of $K$ captions using beam search or diverse decoding strategies. Given $K$ captions, the recall of the concept pairs in an evaluation dataset is:

$$\text{Recall@K} = \frac{|\{(s'_{ik}) \mid \exists k : s'_{ik} \in C\}|}{M}$$

Recall@K is an appropriate metric because the reference captions were produced by annotators who did not need to produce any specific word when describing an image. In addition, the set of captions $C$ is determined with respect to the same synonym sets of the concepts that were used to construct the datasets, and so credit is given for semantically equivalent outputs. More exhaustive approaches to determine semantic equivalence for this metric are left for future work.

4 State-of-the-Art Performance

4.1 Experimental Protocol

Models: We evaluate two image captioning models on the compositional generalization task: Show, Attend and Tell (SAT; Xu et al., 2015) and Bottom-up and Top-down Attention (BUTD; Anderson et al., 2018). For SAT, we use ResNet-152 (He et al., 2016) as an improved image encoder.

Training and Evaluation: The models are trained on the $D_{train}$ datasets, in which groups of concept pairs are held out—see Appendix C for more information. Hyperparameters are set as described in the respective papers. When a model has converged on the $D_{val}$ validation split (as measured in BLEU score), we generate $K$ captions for each image in $D_{eval}$ using beam search. Then, we calculate the Recall@K metric (Eqn. 1, $K=5$) for each concept pair in the evaluation split, as well as the average over all recall scores to report the compositional generalization performance of a model.

We also evaluate the compositional generalization of a BUTD model trained on the full COCO dataset...
training dataset (FULL). In this setting, the model is trained on compositions of the type we seek to evaluate in this task, and thus does not need to generalize to new compositions.

Pretrained Language Representations: The word embeddings of image captioning models are usually learned from scratch, without pretraining. Pretrained word embeddings (e.g. GloVe (Pennington et al., 2014)) or language models (e.g. Devlin et al. (2019)) contain distributional information obtained from large-scale textual resources, which may improve generalization performance. However, we do use them for this task because the resulting model may not have the expected paradigmatic gaps.

4.2 Results

Image Captioning: The models mostly fail to generate captions that contain the held out pairs. The average Recall@5 for SAT and BUTD are 3.0 and 6.5, respectively. A qualitative analysis of the generated captions shows that the models usually describe the depicted objects correctly, but, in the case of held out adjective–noun pairs, the models either avoid using adjectives, or use adjectives that describe a different property of the object in question, e.g. white and green airplane instead of small plane in Figure 3. In the case of held out noun–verb pairs, the models either replace the target verb with a less descriptive phrase, e.g. a man sitting with a plate of food instead of a man is eating in Figure 3, or completely omit the verb, reducing the caption to a simple noun phrase.

In the FULL setting, average Recall@5 reaches 33.3. We assume that this score is a conservative estimate due to the low average inter-annotator agreement (see Footnote 6). The model is less likely to describe an image using the target pair if the pair is only present in one of the reference captions, as the feature is likely not salient (e.g. the car in the image has multiple colors, and the target color is only covering one part of the car). In fact, if we calculate the average recall for images where at least 2 / 3 / 4 / 5 of the reference captions contain the target concept pair, Recall@5 increases to 46.5 / 58.3 / 64.9 / 75.2. This shows that the BUTD model is more likely to generate a caption with the expected concept pair when more human annotators agree that it is a salient pair of concepts in an image.

Image–Sentence Ranking: In a related experiment, we evaluate the generalization performance of the VSE++ image–sentence ranking model on the compositional captioning task (Faghri et al., 2018). We use an adapted version of the evaluation metric because the ranking model does not generate tokens. The average Recall@5 with the adapted metric for the ranking model is 46.3. The respective FULL performance for this model is 47.0, indicating that the model performs well whether it has seen examples of the evaluation concept pair at training time or not. In other words, the model achieves better compositional generalization than the captioning models.

5 Joint Model

In the previous section, we found that state-of-the-art captioning models fail to generalize to unseen combinations of concepts, however, an image-sentence ranking model does generalize. We propose a multi-task model that is trained for image captioning and image–sentence ranking with shared parameters between the different tasks. The captioning component can use the ranking component to re-rank complete candidate captions in the beam. This ensures that the generated captions are as informative and accurate as possible, given the constraints of satisfying both tasks.

Following Anderson et al. (2018), the model is a two-layer LSTM (Hochreiter and Schmidhuber, 1997), where the first layer encodes the sequence of words, and the second layer integrates visual features from the bottom-up and top-down attention mechanism, and generates the output sequence. The parameters of the ranking component $\theta_2$ are mostly a subset of the parameters of the generation component $\theta_1$. We name the model Bottom-Up and Top-down attention with Ranking (BUTR). Figure 2 shows a high-level overview of the model architecture.

5.1 Image–Sentence Ranking

To perform the image–sentence ranking task, we project the images and captions into a joint visual-semantic embedding space $\mathbb{R}^d$. We introduce a
language encoding LSTM with a hidden layer dimension of $L$.

$$h^l_t = \text{LSTM}(W_1 o_t, h^{l-1}_t) \quad (2)$$

where $o_t \in \mathbb{R}^V$ is a one-hot encoding of the input word at timestep $t$, $W_1 \in \mathbb{R}^{E \times V}$ is a word embedding matrix for a vocabulary of size $V$ and $h^{l-1}_t$ the state of the LSTM at the previous timestep. At training time, the input words are the words of the target caption at each timestep.

The final hidden state of the language encoding LSTM $h^l_{t=T}$ is projected into the joint embedding space as $s^* \in \mathbb{R}^J$ using $W_2 \in \mathbb{R}^{J \times L}$:

$$s^* = W_2 h^l_{t=T} \quad (3)$$

The images are represented using the bottom-up features proposed by Anderson et al. (2018). For each image, we extract a set of $R$ mean-pooled convolutional features $v_r \in \mathbb{R}^J$, one for each proposed image region $r$. We introduce $W_3 \in \mathbb{R}^{J \times J}$, which projects the image features of a single region into the joint embedding space:

$$v^e_r = W_3 v_r \quad (4)$$

To form a single representation $v^*$ of the image from the set of embedded image region features $v^e_r$, we apply a weighting mechanism. We generate a normalized weighting of region features $\alpha \in \mathbb{R}^R$ using $W_4 \in \mathbb{R}^{1 \times J}$. $\beta_r$ denotes the weight for a specific region $r$. Then we sum the weighted region features to generate $v^* \in \mathbb{R}^J$:

$$\beta'_r = W_4 v^e_r \quad (5)$$

$$\beta = \text{softmax}(\beta') \quad (6)$$

$$v^* = \sum_{r=1}^R \beta_r v^e_r \quad (7)$$

We define the similarity between an image and a caption as the cosine similarity $\cos(v^*, s^*)$.

5.2 Caption Generation

For caption generation, we introduce a separate language generation LSTM that is stacked on top of the language encoding LSTM. At each timestep $t$, we first calculate a weighted representation of the input image features. We calculate a normalized attention weight $\alpha_t \in \mathbb{R}^R$ (one $\alpha_{r,t}$ for each region) using the language encoding and the image region features. Then, we create a single weighted image feature vector:

$$o_{r,t} = W_5 \text{tanh}(W_6 v^e_r + W_7 h^g_{t-1}) \quad (8)$$

$$\alpha_t = \text{softmax}(\alpha'_{r,t}) \quad (9)$$

$$v_t = \sum_{r=1}^R \alpha_{r,t} v^e_r \quad (10)$$

where $W_5 \in \mathbb{R}^H$, $W_6 \in \mathbb{R}^{H \times J}$ and $W_7 \in \mathbb{R}^{H \times L}$. $H$ indicates the hidden layer dimension of the attention module.

These weighted image features $v_t$, the output of the language encoding LSTM $h^g_t$ (Eqn. 2) and the previous state of the language generation LSTM $h^g_{t-1}$ are input to the language generation LSTM:

$$h^g_t = \text{LSTM}([v_t, h^g_t], h^g_{t-1}) \quad (11)$$

The hidden layer dimension of the LSTM is $G$. The output probability distribution over the vocabulary is calculated using $W_8 \in \mathbb{R}^{V \times G}$:

$$p(w_t | w_{<t}) = \text{softmax}(W_8 h^g_t) \quad (12)$$

5.3 Training

The model is jointly trained on two objectives. The caption generation component is trained with
a cross-entropy loss, given a target ground-truth sentence $s$ consisting of the words $w_1, \ldots, w_T$:

$$\mathcal{L}_{\text{gen}}(\theta_1) = -\sum_{t=1}^{T} \log p(w_t|w_{<t}; \theta)$$  \hspace{1cm} (13)$$

The image-caption ranking component is trained using a hinge loss with emphasis on hard negatives (Faghri et al., 2018):

$$\mathcal{L}_{\text{rank}}(\theta_2) = \max_{s'}[\alpha + \cos(i, s') - \cos(i, s)] +$$

$$+ \max_{i'}[\alpha + \cos(i', s) - \cos(i, s)]$$  \hspace{1cm} (14)$$

where $[x]_+ \equiv \max(x, 0)$.

These two loss terms can take very different magnitudes during training, and thus can not be simply added. We use GradNorm (Chen et al., 2018) to learn loss weighting parameters $w_{\text{gen}}$ and $w_{\text{rank}}$ with an additional optimizer during training. These parameters dynamically rescale the gradients so that no task becomes too dominant. The overall training objective is formulated as the weighted sum of the single-task losses:

$$\mathcal{L}(\theta_1, \theta_2) = w_{\text{gen}}\mathcal{L}_{\text{gen}}(\theta_1) + w_{\text{rank}}\mathcal{L}_{\text{rank}}(\theta_2)$$  \hspace{1cm} (15)$$

### 5.4 Inference

The model generates $B$ captions for each image using beam search decoding. At each timestep, the tokens generated so far for each item on the beam are input back into the language encoder (Eqn. 3). The output of the language encoder is concatenated with the image representation (Eqn. 7) and the previous hidden state of the generation LSTM, and input to the generation LSTM (Eqn. 11) to predict the next token (Eqn. 12).

The jointly-trained image–sentence ranking component can be used to re-rank the generated captions comparing the image embedding with a language encoder embedding of the captions (Eqn. 4). We expect the ranking model will produce a better ranking of the $B$ captions than only beam search by considering their relevance and informativity with respect to the image.

### 6 Results

We follow the experimental protocol defined in Section 4 to evaluate the joint model. See Appendix E for training details and hyperparameters.

Table 2 shows the compositional generalization performance, as well as the common image captioning metric scores for all models. **BUTR** uses the same image features and a decoder architecture as the **BUTD** model. Thus, when using the standard beam search decoding method, **BUTR** does not improve over **BUTD**. However, when using the improved decoding mechanism with re-ranking **BUTR + RR**, **Recall@5** increases to 13.2. We also observe an improvement in **METEOR** and **SPICE**, and a drop in **BLEU** and **CIDEr** compared to the other models. We note that **BLEU** has the weakest correlations (Elliott and Keller, 2014), and **SPICE** and **METEOR** have the strongest correlations with human judgments (Kilickaya et al., 2017).

The **Recall@5** scores for different categories of held out pairs is presented in Table 3, and Figure 3 presents examples of images and the generated captions from different models. We observe that all models are generally best at describing colors, especially of inanimate objects; they nearly never correctly describe held out size modifiers; and for held out noun–verb pairs, performance is slightly better for transitive verbs.

### Table 2: Average results for **Recall@5** (R; Eqn. 1), **METEOR** (M; Denkowski and Lavie, 2014), **SPICE** (S; Anderson et al., 2016), **CIDEr** (C; Vedantam et al., 2015), **BLEU** (B; Papineni et al., 2002). **RR** stands for re-ranking after decoding.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>M</th>
<th>S</th>
<th>C</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT</td>
<td>3.0</td>
<td>23.2</td>
<td>16.6</td>
<td>80.4</td>
<td>27.5</td>
</tr>
<tr>
<td>BUTD</td>
<td>6.5</td>
<td>25.8</td>
<td>19.1</td>
<td>98.1</td>
<td>32.6</td>
</tr>
<tr>
<td>BUTR</td>
<td>6.5</td>
<td>25.7</td>
<td>19.0</td>
<td>97.0</td>
<td>32.0</td>
</tr>
<tr>
<td><strong>BUTR + RR</strong></td>
<td><strong>13.2</strong></td>
<td><strong>26.4</strong></td>
<td><strong>20.4</strong></td>
<td>92.7</td>
<td>28.8</td>
</tr>
<tr>
<td>FULL</td>
<td>33.3</td>
<td>27.4</td>
<td>20.9</td>
<td>105.3</td>
<td>36.6</td>
</tr>
</tbody>
</table>

### Table 3: Detailed **Recall@5** scores for different categories of held out pairs. The scores are averaged over the set of scores for pairs from the respective category. **RR** stands for re-ranking after decoding. Color and size adjectives are split into **Animate** or **Inanimate** objects; Verbs are split into **Transitive** and **Intransitive** verbs.

<table>
<thead>
<tr>
<th></th>
<th>Color</th>
<th>Size</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>I</td>
<td>A</td>
</tr>
<tr>
<td>SAT</td>
<td>3.7</td>
<td>10.5</td>
<td>0</td>
</tr>
<tr>
<td>BUTD</td>
<td>5.4</td>
<td>10.9</td>
<td>0.5</td>
</tr>
<tr>
<td>BUTR</td>
<td>6.4</td>
<td>16.2</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>+ RR</strong></td>
<td><strong>13.8</strong></td>
<td><strong>26.0</strong></td>
<td><strong>1.4</strong></td>
</tr>
<tr>
<td>FULL</td>
<td>42.7</td>
<td>38.7</td>
<td>5.9</td>
</tr>
</tbody>
</table>
7 Analysis and Discussion

**Describing colors:** The color–noun pairings studied in this work have the best generalization performance. We find that all models are better at generalizing to describing inanimate objects instead of animate objects, as shown in the detailed results in Table 3. One explanation for this could be that the colors of inanimate objects tend to have a higher variance in chromaticity when compared to the colors of animate objects (Rosenthal et al., 2018), making them easier to distinguish.

**Describing sizes:** The generalization performance for size modifiers is consistently low for all models. The CNN image encoders are generally able to predict the sizes of object bounding boxes in an image. However, this does not necessarily relate to the actual sizes of the objects, given that this depends on their distance from the camera. To support this claim, we perform a correlation analysis in Appendix F showing that the bounding box sizes of objects in the COCO dataset do not relate to the described sizes in the respective captions.

Nevertheless, size modification is challenging from a linguistic perspective because it requires reference to an object’s comparison class (Cresswell, 1977; Bierwisch, 1989). A large mouse is so with respect to the class of mice, not with respect to the broader class of animals. To successfully learn size modification, a model needs to represent such comparison classes.

We hypothesize that recall is reasonable in the FULL setting because it exploits biases in the dataset, e.g. that trucks are often described as BIG.

In that case, the model is not actually learning the meaning of BIG, but simple co-occurrence statistics for adjectives with nouns in the dataset.

**Describing actions:** In these experiments, the models were better at generalizing to transitive verbs than intransitive verbs. This may be because images depicting transitive events (e.g. eating) often contain additional arguments (e.g. cake); thus they offer richer contextual cues than images with intransitive events. The analysis in Appendix G provides some support for this hypothesis.

**Diversity in Generated Captions:** A crucial difference between human-written and model-generated captions is that the latter are less diverse (Devlin et al., 2015; Dai et al., 2017). Given that BUTR+RR improves compositional generalization, we explore whether the diversity of the captions is also improved. Van Miltenburg et al. (2018) proposes a suite of metrics to measure the diversity of the captions generated by a model. We apply these metrics to the captions generated by BUTR+RR and BUTD and compare the scores to the best models evaluated in Van Miltenburg et al. (2018).

The results are presented in Table 4. BUTR+RR shows the best performance as measured by most of the diversity metrics. BUTR+RR produces the highest percentage of novel captions (%Novel), which is important for compositional generalization. It generates sentences with a high average sentence length (ASL) – performing similarly to Liu et al. (2017) – but with a larger standard deviation, suggesting a greater variety in the captions. The total number of word types (Types) and cover-
age (Cov) are higher for Shetty et al. (2017), which is trained with a generative adversarial objective in order to generate more diverse captions. However, these types are more equally distributed in the captions generated by BUTR+RR, as shown by the higher mean segmented type-token ratio (TTR₁) and bigram type-token ratio (TTR₂).

The increased diversity of the captions may explain the lower BLEU score of BUTR+RR compared to BUTD. Recall that BLEU measures weighted n-gram precision, hence it awards less credit for captions that are lexically or syntactically different than the references. Thus, BLEU score may decrease if a model generates diverse captions. We note that METEOR, which incorporates non-lexical matching components in its scoring function, is higher for BUTR+RR than BUTD.

Decoding strategies: The failure of the captioning models to generalize can be partially ascribed to the effects of maximum likelihood decoding. Holtzman et al. (2019) find that maximum likelihood decoding leads to unnaturally flat and high per-token probability text. We find that even with grounding from the images, the captioning models do not assign a high probability to the sequences containing compositions that were not observed during training. BUTR is jointly trained with a ranking component, which is used to re-rank the generated captions, thereby ensuring that at the sentence-level, the captions are relevant for the image. It can thus be viewed as an improved decoding strategy such as those proposed in Vijayakumar et al. (2018); Fan et al. (2018); Radford et al. (2019); Holtzman et al. (2019).

8 Conclusion

Image captioning models are usually evaluated without explicitly considering their ability to generalize to unseen concepts. In this paper, we argued that models should be capable of compositional generalization, i.e. the ability to produce captions that include combinations of unseen concepts. We evaluated the ability of models to generalize to unseen adjective–noun and noun–verb pairs and found that two state-of-the-art models did not generalize in this evaluation, but that an image–sentence ranking model did. Given these findings, we presented a multi-task model that combines captioning and image–sentence ranking, and uses the ranking component to re-rank the captions generated by the captioning component. This model substantially improved generalization performance without sacrificing performance on established text-similarity metrics, while generating more diverse captions. We hope that this work will encourage researchers to design models that better reflect human-like language production.

Future work includes extending the evaluation to other concept pairs and other concept classes, analysing the circumstances in which the re-ranking step improves compositional generalization, exploring the utility of jointly trained discriminative re-rankers into other NLP tasks, developing models that generalize to size modifier adjectives, and devising approaches to improve the handling of semantically equivalent outputs for the proposed evaluation metric.

Acknowledgements

We thank Emiel van Miltenburg and Ákos Kádár for their extensive feedback on the work, and the reviewers, Ana Valeria Gonzales, Daniel Hershcovitch, Heather Lent, and Mareïke Hartmann for their comments. We also thank the participants of the Lorentz Center workshop on Compositionality in Brains and Machines for suggesting the phrase “paradigmatic gap”. MN was supported by the Erasmus+ Traineeship program. RA and MA are funded by a Google Focused Research Award.
References


Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In \textit{The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)}.


