Retrieval, Analogy, and Composition: A framework for Compositional Generalization in Image Captioning

Zhan Shi*, **Hui Liu***, **Martin Renqiang Min**†, **Christopher Malon**†
Li Erran Li‡, **Xiaodan Zhu***
*Ingenuity Labs Research Institute & ECE, Queen's University
†NEC Laboratories America
‡AWS AI, Amazon
{z.shi, hui.liu, xiaodan.zhu}@queensu.ca
{renqiang, malon}@nec-labs.com, erranli@gmail.com

Abstract

Image captioning systems are expected to have the ability to combine individual concepts when describing scenes with concept combinations that are not observed during training. In spite of significant progress in image captioning with the help of the autoregressive generation framework, current approaches fail to generalize well to novel concept combinations. We propose a new framework that revolves around probing several similar image caption training instances (retrieval), performing analogical reasoning over relevant entities in retrieved prototypes (analogy), and enhancing the generation process with reasoning outcomes (composition). Our method augments the generation model by referring to the neighboring instances in the training set to produce novel concept combinations in generated captions. We perform experiments on the widely used image captioning benchmarks. The proposed models achieve substantial improvement over the compared baselines on both composition related evaluation metrics and conventional image captioning metrics.

1 Introduction

Generating a textual description for a given image, a problem known as image captioning (Chen et al., 2015), requires a conditional generation model to recognize salient visual regions, e.g., object (Anderson et al., 2018) or scene graph detection (Yao et al., 2018), align visual features with textual tokens (Lu et al., 2017; Pu et al., 2018; Shi et al., 2020), and verbalize them in a natural language sentence (Xu et al., 2015; Lu et al., 2018). Current state-of-the-art image captioning models benefit from powerful neural autoregressive generation models, attention mechanisms, and progress in object or scene graph detection. They have achieved significant progress in obtaining visual representations for images as well as modelling alignment between visual features and textual tokens, resulting in superior performance under a variety of text-similarity based metrics.

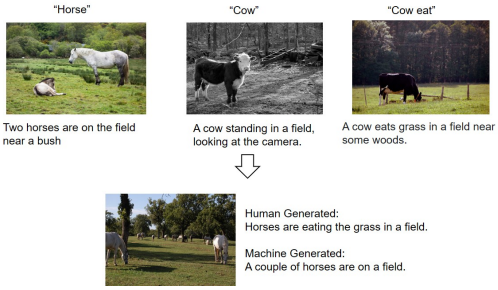

Figure 1: Comparison of compositional generalization in generated descriptions between human and machine (Anderson et al., 2018).

However, when verbalising the visual semantic concepts into natural language sentences, these models still fall short of compositional generalization for images with novel concept combinations (Nikolaus et al., 2019). Note that making systematic generalizations (Lake and Baroni, 2018; Janssen and Partee, 1997) from limited data is an essential property of human language. As shown in Figure 1, the visual instances of *"horse"* and *"cow"* as well as the scene containing concept combinations of *"cow eat"* have been observed during training. While the existing models can often generate *"horse on"* for the picture, it would be effortless for humans to generate a caption containing *"horse eat"* even this combination has not been observed during training. It is partly due to the fact that current language generation models rely heavily on the surface distributional characteristics of the captions and hence are discouraged from generating unseen concept combinations (Holtzman et al., 2019; Nikolaus et al., 2019).

To remedy the problem, we propose to leverage prototype-based generation approaches (Guu et al., 2018; Hashimoto et al., 2018) which can explicitly expose concepts of other training examples by asking the model to decide what prototypes to retrieve in either a heuristic or learned way. In other

words, these approaches have a chance to peek into retrieved prototypes for concepts without relying on the generation component. In addition, to combine the concepts from the prototypes, we enhance the conditional generation model by incorporating analogical reasoning (Vosniadou and Ortony, 1989; Gentner and Smith, 2012; Wu et al., 2020), based on the idea that, if two things are similar on the visual side, they are probably also similar on the text side. Specifically, in each generation step, we compare the visual and textual representation between the current state in the language model decoder and analogy entity pairs (a visual entity and its text form a pair) extracted from retrieved prototypes to produce sentences with improved generalization of semantic compositions.

As a result, our model consists of two major components: (1) a multi-prototype retriever (c.f. Section 3.3) for obtaining multiple prototypes, which aims to cover the basic concepts in the described image, and (2) an analogical reasoning editor (c.f. Section 3.4) to perform analogical reasoning over extracted analogy entity pairs, in order to compose these concepts for generation. We perform extensive experiments on the widely used benchmark MSCOCO (Lin et al., 2014) with both maximum likelihood estimation and reinforcement learning strategies (Rennie et al., 2017). The experiment results show that the proposed models significantly outperform the baselines under both text-similarity based metrics and composition related metrics. The main contributions of our work are summarized as follows:

- To the best of our knowledge, this is the first attempt to introduce a novel prototype-based generation framework in image captioning, which helps the generation process with improved compositional generalization.
- The proposed framework substantially improves upon the baselines (Anderson et al., 2018; Nikolaus et al., 2019) on both composition related metrics (from 13.6 to 18.8 on R@5) and conventional evaluation metrics (from 109.9 to 114.3 on CIDEr).
- We analyze various types of concept composition in captioning generation and provide detailed discussion on how the proposed framework improves compositional generalization for each type.

2 Related Work

Image Caption Generation Image captioning aims at generating visually grounded descriptions for images. Current models often leverage a CNN or variants as the image encoder and an RNN or transformer as the decoder to generate sentences (Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Donahue et al., 2015; Yang et al., 2016; Huang et al., 2019). Previous work has used a visual attention mechanism (Anderson et al., 2018; Pu et al., 2018; Lu et al., 2017; Pedersoli et al., 2017; Xu et al., 2015; Pan et al., 2020; Shi et al., 2021b), explicit high-level attributes detection (Yao et al., 2017; Wu et al., 2016; You et al., 2016) to align visual and textual features. For the learning method, people use reinforcement learning methods (Rennie et al., 2017; Ranzato et al., 2015; Liu et al., 2018), or contrastive or adversarial learning (Dai and Lin, 2017; Dai et al., 2017) to generate descriptive captions (Luo et al., 2018; Shi et al., 2021a) with improved quality. The distribution shift between training and test stages also has received a lot of attention, such as generating captions with novel concepts (Lu et al., 2018; Agrawal et al., 2019; Anderson et al., 2016a). More recently, Nikolaus et al. (2019) proposes 24 concept pairs to explicitly investigate the composition generation ability of current neural image captioning models.

Compositional generalization Systematic compositionality, a method to capture underlying rules from limited data and generalize them to novel situations, is a key feature in human intelligence (Fodor and Pylyshyn, 1988). The topic is closely related to cognitive science (Fodor and Lepore, 2002) and connectionist literature (McClelland et al., 1986). While the topic is widely studied in the semantic parsing literature (Lake and Baroni, 2018; Keysers et al., 2019), it is less investigated in natural language generation. Akyürek et al. (2020) introduces a resample and recombine network to improve generalization in two NLP problems, i.e., instruction following and morphological analysis.

3 Method

Our framework is designed to enhance text generation with compositional generalization through analogical reasoning from retrieved prototypes. The framework is built on the classical two-layer LSTM network, i.e., Updown (Anderson et al., 2018), but this method is orthogonal to more re-

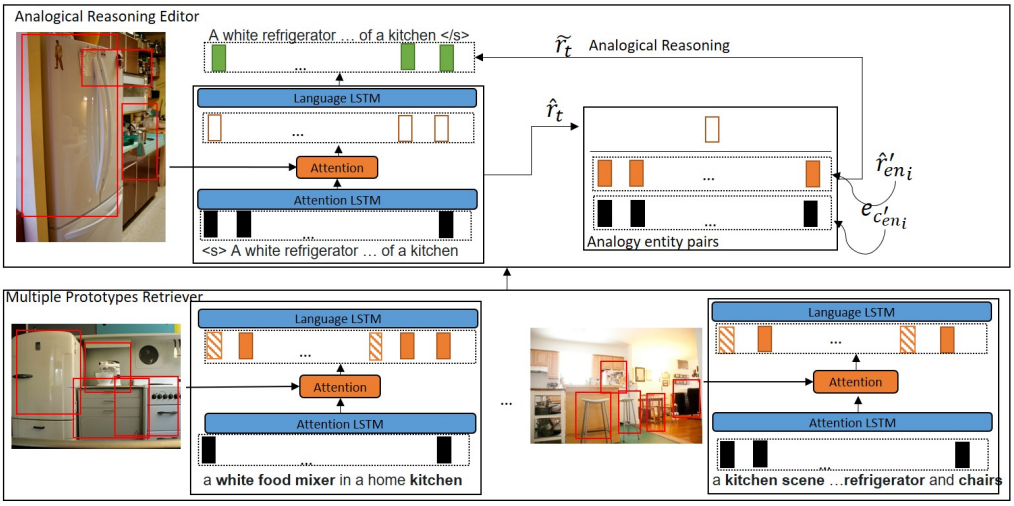

Figure 2: The model framework consists of a prototype retriever and analogical reasoning editor where the former attempts to obtain multiple prototypes to the described image and the latter uses analogical reasoning to leverage the analogy entity pairs for generation. Therefore, even if "white refrigerator" is a novel combination, we can generate a captioning containing it from entities in prototypes.

cent transformer based image captioning networks, e.g., AoANet (Huang et al., 2019) and M2 Transformer (Cornia et al., 2020). We will leave combining these ideas for future work.

3.1 Problem Definition

We are given a training dataset D which contains matched image-caption pairs $\{d_i\}$, where d_i denotes an image x_i and its caption c_i.

Composition of Common Concepts Following (Nikolaus et al., 2019), we use some common concepts $\{s_i\}$ of interest, which covers a range of different types of attributes, objects and verbs frequently and then select a number of concept pairs $\{S_j\}$ based on $\{s_i\}$, including attribute-noun and noun-verb composition. Note that $\{s_i\}$ is frequently seen in both training and evaluation stages but $\{S_j\}$ is a held-out set of concept combinations to test the generalization ability of the model. (c.f. Section 4.1 for dataset splits).

Composition of Rare Concepts We further select a few rare concepts $\{s'_i\}$ of interest, covering a few verbs and objects. As these concepts are rarely seen in the training stage but frequently used in evaluation stage, these rare concepts are proposed to test the generalization ability to learn new concepts in context from little data. (c.f. Section 4.1 for dataset splits)

3.2 Overall Framework

The goal of image captioning is to train a conditional generation model $p_m(c \mid x)$. As shown in Figure 2, the framework corresponds to the following **retrieve** and **edit** generative process: given an input x, we first retrieve k prototypes $d'_{1:k}$ from D by sampling from $p_r(d'_{1:k} \mid x)$. We then generate a visually grounded sentence c using an analogical reasoning editor $p_e(c \mid x, d'_{1:k})$.

Typical models leverage a two-phase training process to learn $p_m(c \mid x)$: the former phase uses the cross entropy loss to maximize the log probability with respect to the ground truth captions and the latter phase uses a policy gradient algorithm to maximize the expected reward metric r, i.e., CIDEr:

$$p_m(c \mid x) = \sum_{d'_{1:k}} p_e(c \mid x, d'_{1:k}) p_r(d'_{1:k} \mid x) \quad (1)$$

$$\mathcal{L}_{CE} = E_{(x,c)\sim D}[\log p_m(c \mid x)] \quad (2)$$

$$\mathcal{L}_{RL} = E_{\hat{c}\sim p_m(c|x)}[r(\hat{c}, x)] \quad (3)$$

Here, we focus on deterministic retrievers, where $p_r(d'_{1:k} \mid x)$ is a point mass on particular prototypes $d'_{1:k}$.

Note that when generating texts with novel semantic compositions, neither a basic LSTM editor p_e nor a single prototype retriever p_r is enough. We accordingly elaborate on retrieval and edit models separately in the rest of this section.

3.3 Multi-Prototype Retriever

To generate captions with novel compositions, we aim to provide a large inventory of contextualized individual concepts and encourage further use of both their visual and textual features. Furthermore,

the retrieved prototypes not only focus on visual similarity with the query, but also gather enough information to cover the concepts in the query collectively.

Specifically, given an image x, we first obtain n neighbor prototypes $x_{1:n}$ in the training set by ranking the cosine similarity of image features encoded by CLIP (Radford et al., 2021), which is the state-of-the-art visual encoder trained on a large amount of image-text data by contrastive learning. Then for each neighbor image x_i and the query image x, we get its entity[1] set g_{x_i} and g_x from the scene graph using a pre-trained parser (Yang et al., 2019). We get K images $\{x'_j\}_{j=1}^{K}$ by iteratively selecting from the following formulas:

$$x'_{j+1} = \arg\max_{x_i \in x_{1:n}} \frac{|g_{x_i} \cap g_{sub}|}{|g_{x_i}|} \tag{4}$$

$$g_{sub} = \begin{cases} g_x - \cup_{m=1}^{j} g_{x'_m} & j > 0 \\ g_x & j = 0 \end{cases} \tag{5}$$

As such we obtain K retrieved images with their corresponding captions $d'_{1:K}$ so that these K prototypes cover most meaningful semantic concepts in the query image x.

3.4 Analogical Reasoning Editor

We take $p_e(c \mid d'_{1:k}, x)$ to be a neural autoregressive conditional text generation model (a two layer LSTM) which decomposes as: $p_e(c \mid d'_{1:k}, x) = \prod_{t=1}^{T} p_e(c_t \mid c_{<t}, d'_{1:k}, x)$, where T is the length of the caption and c_0 is the start token "<s>". For image x the model employs Faster R-CNN (Ren et al., 2015) to recognize instances of objects and returns a set of image regions for objects: $x = \{r_1, r_2, \cdots, r_M\}$.

Bottom LSTM The bottom LSTM is used to align a textual state to image region representations:

$$\mathbf{h}_t^1 = \text{LSTM}(\mathbf{h}_{t-1}^1, [\mathbf{h}_{t-1}^2; \overline{r}; \mathbf{e}_{c_{t-1}}]) \tag{6}$$

where LSTM means one step of recurrent unit computation via LSTM; $\overline{r}$ is the mean-pooled representation of all object regions in the image; $\mathbf{h}_{t-1}^1$ and $\mathbf{h}_{t-1}^2$ denote hidden states of bottom and top LSTM at time step $t-1$, respectively; $\mathbf{e}$ is the word embedding lookup table.

[1]Entity means attributes, objects and predicates here

Attention Unit The state $\mathbf{h}_t^1$ is then used as a query to attend over object features $\{r_i\}$ to get contextualized image region features $\hat{r}_t$:

$$a_{i,t} = W_a^T \tanh(W_{ra} r_i + W_{ha} \mathbf{h}_t^1) \tag{7}$$

$$\alpha_t = \text{softmax}(a_t) \tag{8}$$

$$\hat{r}_t = \sum_{i=1}^{M} \alpha_{i,t} r_i \tag{9}$$

where W_{ra}, W_{ha} and W_a are model parameters.

Top LSTM The top-layer LSTM works as a recurrent language model. At time step t, the input consists of the output from the bottom LSTM layer $\mathbf{h}_t^1$ and the output of visual attention unit $\hat{r}_t$:

$$\mathbf{h}_t^2 = \text{LSTM}(\mathbf{h}_{t-1}^2, [\mathbf{h}_t^1; \hat{r}_t]) \tag{10}$$

Analogy Entity Pairs We first run the two-layer LSTM on the K retrieved prototypes $d'_{1:K}$ to obtain aligned visual and textual representations. We take the attention unit outcome as the visual feature and its corresponding ground truth token as the textual feature, obtaining a total of $K \cdot T$ aligned pairs. Specifically, in time step t and retrieved prototype k, we get the aligned pair as $\{(\mathbf{e}_{c'_{k,t}}, \hat{r}'_{k,t})\}$. To obtain the analogy entity pairs, we remove the pair if $c'_{k,t}$ is not an entity, thus getting Y (Y is dependent on the input x and its retrieved prototype $d'_{1:K}$) analogy entity pairs $\{(\mathbf{e}_{c'_{en_i}} \hat{r}'_{en_i})\}, 1 \le i \le Y$.

Analogical Reasoning For the described image x, we obtain the analogy entity pairs $\{(\mathbf{e}_{c'_{en_i}}, \hat{r}'_{en_i})\}$. An analogy pair consists of a pair of visual and textual features, and analogical reasoning is the type of reasoning that relies upon the analogy pairs. we perform analogical reasoning over these analogy pairs. Specifically, we use $\hat{r}_t$ as the query for attending these entity pairs to get analogy context features:

$$b_{i,t} = W_b^T \tanh(W_{rb} \hat{r}_t + W_{hb} \hat{r}'_{en_i}) \tag{11}$$

$$\beta_t = \text{softmax}(b_t) \tag{12}$$

$$\tilde{r}_t = \sum_{i=1}^{Y} \beta_{i,t} \mathbf{e}_{en_i} \tag{13}$$

We combine features from $\tilde{r}_t$ and the top layer LSTM hidden state $\mathbf{h}_t^2$ to predict the next token:

$$p_e(c_t \mid c_{<t}, d'_{1:k}, x) = \text{softmax}(W_p[\mathbf{h}_t^2; \tilde{r}_t] + b_p) \tag{14}$$

4 Experiments

4.1 Datasets and Experiment Setup

MSCOCO We perform extensive experiments on the MSCOCO benchmark (Lin et al., 2014). Corresponding to the generalization of both common and rare concepts, we handcraft a few data splits for training and evaluation. (a) Common concepts: Nikolaus et al. (2019) selects 12 nouns, 7 attributes and 6 verbs as common concepts and 24 concept combinations as a held-out set. They build four different train/val/test splits with 6 concept combinations as a group. The four groups are i: (black_cat, big_bird, red_bus, small_plane, man_eat, woman_lie); ii: (brown_dog, small_cat, white_truck, big_plane, woman_ride, bird_fly); iii: (white_horse, big_cat, blue_bus, small_table, child_hold, bird_stand); iv: (black_bird, small_dog, white_boat, big_truck, horse_eat, child_stand). (b) Rare concepts: We select 3 nouns and 3 verbs as rare concepts. We build one split to the above 6 concepts. (c.f. Section 4.3 for the split construction process). Table 1 shows specific common concept, rare concept, and Karparthy split (Karpathy and Fei-Fei, 2015) information.

Split	Train	Val	Test	Held-out
Common	79K	3K	1K	Group i
	79K	3K	1K	Group ii
	79K	3K	1K	Group iii
	79K	3K	1K	Group iv
Rare	72K	10K	5K	horse, bench, sleep, smile, jump ,plane
Karparthy	113K	5K	5K	N/A

Table 1: Statistics of different splits.

4.2 Evaluation Metrics

Quality-related We employ a wide range of conventional reference based image caption evaluation metrics, i.e., SPICE(SP) (Anderson et al., 2016b), CIDEr(CD) (Vedantam et al., 2015), METEOR(ME) (Denkowski and Lavie, 2014), ROUGE-L(RG) (Lin, 2004), and BLEU (Papineni et al., 2002) to evaluate the generated captions.

Diversity-related We report diversity by calculating the number of distinctly generated unigrams(Div-1) and bigrams(Div-2) scaled by sentence length (Li et al., 2015) as well as self-BLEU (Zhu et al., 2018) (a lower value yields a higher diversity), which is computed among multiple generated sentences.

Composition-related We calculate the recall of the concept pairs (R@K) (Nikolaus et al., 2019) for the multiple (K) generated captions given images in the evaluation dataset.

4.3 Implementation Details

Parameter Setting To make a fair comparison, we use the default experiment setup that the compared baselines used as indicated in Luo's package[2]. The number of retrieved prototypes k is 3 and the specific retrieval model used for obtaining prototypes is ViT-B/32 by official release (note that in prototype retrieval, we only use the image encoder). The leveraged scene graph parser is the same with the official release from (Yang et al., 2019). For the decoding stage, we use beam search to produce 5 sentences, i.e., the beam size is also 5, for further evaluation. The re-rank strategy is based on a beam search with size of 100, and then ranking the sentences in the beam by the ViT-B/32.

Split Construction We first use a set of synonyms (Nikolaus et al., 2019) to represent one concept as each concept accounts for the variations it can be expressed across the dataset. Then we use the dependency parser from StanfordNLP (Qi et al., 2019) to identify the chosen nouns, verbs, attributes, noun-verb, and attribute-noun concept combinations. For the construction of rare concept splits, we pick up all image-caption pairs in the original training set that contain the rare concept and distribute 95 percent of them into the validation set, leaving 5 percent of the pairs unchanged in the training set.

4.4 Quantitative Analysis

4.4.1 Overall Performance

Composition and diversity related metrics. We analyze the composition and diversity related metrics together to have a clearer view of compositional generalization ability, as intuitively a more diversified generation method would be helpful in increasing the R@5 of generating concept pairs in the sentence. As shown in Table 2:

(a) On the common concept split, our method achieves a significant increase of compositional generalization, improving the recall@5 from 7.0

[2]https://github.com/ruotianluo/ImageCaptioning.pytorch

	Common Concept Split							Rare Concept Split						
	R@5	ME	CD	SP	Div1	Div2	sB4	R@5	ME	CD	SP	Div1	Div2	sB4
UD (Anderson et al., 2018)	7.0	27.7	99.4	19.9	27.2	35.7	81.2	13.5	26.2	85.4	18.3	25.9	35.1	80.3
UR (Nikolaus et al., 2019)	7.0	27.6	98.7	19.7	28.1	36.1	80.5	13.3	26.0	84.6	18.1	26.1	35.3	79.6
UR+Rank (Nikolaus et al., 2019)	13.6	**28.2**	92.6	20.3	33.2	44.9	62.6	15.8	26.7	81.8	19.1	32.2	44.1	63.8
Ours	10.3	27.2	**101.3**	20.2	26.8	34.9	81.0	15.5	26.3	**88.1**	18.4	25.6	34.7	79.9
Ours+Rank	**18.8**	28.0	94.5	**20.6**	34.8	45.1	61.8	**18.7**	**26.7**	83.2	**19.8**	32.6	44.3	63.9

Table 2: Model performances on the MSCOCO dataset in both common concept and rare concept splits.

	Karparthy split					
	B@1	B@4	ME	RG	CD	SP
UD	75.0	35.0	27.4	55.9	109.9	19.9
Ours	77.3	35.4	26.7	56.8	114.3	20.3
UD-RL	**80.0**	37.2	28.0	57.8	123.5	21.4
Ours-RL	79.5	**37.4**	**28.1**	**57.9**	**125.3**	**21.5**

Table 3: Quality related performances on the Karparthy Split.

	Color		Size		Verb	
	A	I	A	I	T	I
UD	4.4	9.6	0.1	0	13.2	14.7
UR	6.6	15.2	0.1	0.2	7.5	8.6
UR+Rank	11.8	20.6	**2.2**	0.8	30.1	18.2
Ours	4.3	14.6	0.1	0	22.4	19.9
Ours+Rank	**22.0**	**26.3**	1.6	**0.9**	**36.4**	**25.8**

Table 4: Averaged R@5 scores on common concept split. Objects are split into Animate or Inanimate for attributes; Verbs are split into Transitive and Intransitive verbs.

(UD) to 10.3 (Ours) and 13.6 (UR+Rank) to 18.8 (Ours+Rank) with the re-ranking strategy applied.
(b) On the rare concept split, we obtain a similar relative result, increasing recall@5 from 13.5 (UD) to 15.5 (Ours) and 15.8 (UR+Rank) to 18.7 (Ours+Rank) with re-ranking applied.
(c) We can see that the increase of the recall value is not caused by a change of diversity, i.e., the diversity on the common concept split stays almost unchanged from 27.2 (UD) to 26.8 (Ours) in Div1, and 25.9 (UD) to 25.6 (Ours) for Div1 on the rare concept split. However, the re-rank strategy will significantly increase the diversity while improving the recall@5 value.

Table 4 shows more detailed results in terms of various concept combinations. We can see that the increase of performance mostly rests on the noun-verb type concept combinations, increasing from 13.2 (UD) to 22.4 (Ours) and from 14.7 (UD) to 19.9 (Ours) for transitive verbs, i.e., eat, ride, hold, and intransitive verbs, i.e., lie, fly, stand. One explanation for that increase could be derived from the characteristic of prototype retriever, as the retriever is more capable of obtaining prototypes which have similar verbs or nouns with the query image. However, the attribute-noun pairs with size modifiers (big, small) remain the hardest composition generalization problems.

Quality related metrics. The quality-related results from the common concept and rare concept splits, in Table 2, show that our method gains improvement in terms of CIDEr and SPICE, improving CIDEr from 99.4 to 101.3 and SPICE from 19.9 to 20.2. To further verify to what extent the model can improve caption quality, we also test the quality-related metrics on the widely applied Karparthy split. As shown in Table 3, our method consistently outperforms the baseline models on most conventional metrics, especially SPICE and CIDEr in both the CE and RL phases; e.g., the proposed model improves the baseline from 109.9 to 114.3 on CIDEr and 19.9 to 20.3 on SPICE in the CE phase, and 123.5 to 125.3 on CIDEr and 21.4 to 21.5 on SPICE in the RL phase. It is partly due to the fact that this framework can also be viewed as a general method to leverage neighbor instances into training. In contrast to the baseline method that could only condition on the image features for captions, our method can refer to both visual and textual features of multiple prototypes for generation, thus making the models refer to more training examples during inference.

4.4.2 Ablation Analysis

Effect of multiple prototype retriever We analyze the effect of the retriever with regard to the recall value under two aspects: (1) How many prototypes for usage? (2) What kind of retrieved samples would benefit?

Change of prototype numbers We compare recall@5 by changing the prototype numbers in both training and inference stage. As shown in Figure 3, we attempt to use a different number of retrieved

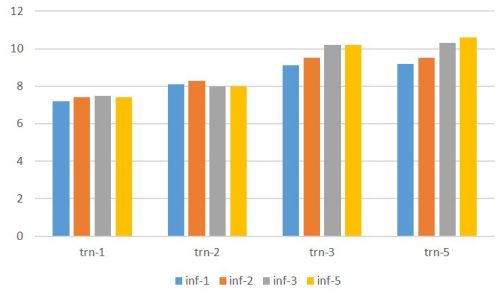

Figure 3: Comparison of use various numbers of prototypes in training and inference. trn-k and inf-k denote use k image in training and inference respectively

Figure 4: Results of different retrievers in training and inference on the common concept split. trn-random retriever and inf-random retriever denote the use of random retriever in both training and inference.

prototypes in training and inference. It shows that the compositional generation ability could be improved with the increasing number of prototypes. The performance gain is marginal when we change the prototype number from 3 to 5. In addition, the model achieves the best performance using the same number of prototypes in training and inference for prototype number of 2, 3 and 5. Using more prototypes in inference would not help for better recall performance.

Change of prototype retrievers To evaluate how the retrievers would affect the recall@5 of concept pairs, we compare two retrievers as below: a random prototype retriever and the retriever used in this work on the common concept split. Note that a random retriever would randomly pick up three image-caption pairs in the training set as prototypes. As shown in Figure 4, we find that using a random retriever in both the training and inference stages would have little improvement over baselines. It demonstrates that the analogy entity pairs extracted from retrieved prototypes play an important role in improving recall@5.

Comparison between CLIP and VSE We also train a visual semantic embedding model (Faghri et al., 2017). Table 5 shows the hit rate of prototypes retrieved by different cross modal retrieval models (The VSE model is trained on the training set of relevant split), e.g., for images containing "black cat" (448), the three prototypes from CLIP can cover "cat" in 420 out of the 448 images, and the three prototypes from VSE can cover "cat" in 405 out of the 448. Overall, we can see from the table that CLIP model shows a better retrieval capacity compared to VSE, achieving a better combination hit in 5 out of the 6 concept pairs. Though both models show similar retrieval performance with regard to nouns, CLIP could yield better performance regarding attributes or verbs.

	Noun		Other		Combine		Total
	C	V	C	V	C	V	
Black cat	420	405	210	160	195	141	448
Big bird	94	94	56	47	40	30	123
Red bus	202	207	151	119	137	103	232
Small plane	149	145	25	16	21	14	158
Man eat	220	233	160	158	134	140	250
Woman lie	121	104	101	87	88	56	144

Table 5: Concept hit of the prototypes between CLIP (C) and VSE (V); Other means verb or attribute

Effect of Analogical Reasoning We analyze the effect of using analogical reasoning over prototype entity pairs compared to the method of mean-pooling the entity pairs representations as the input to the editor. The result shows that the recall value would drop from 10.3 to 7.2 when mean pooling is used, which is almost the same as the baseline (7.0). It demonstrates that aligning the visual features of the described image with the visual features of entity pairs is of critical importance for recall@5.

4.5 Qualitative Analysis

Case Study We list a few cases to show how our model achieves better generation results by the retrieved prototypes. As shown in Figure 5, the image in the first example can retrieve prototypes with similar "red" objects (red lights) and "bus" objects (trolley) and then generate a caption covering the concept of "red bus". For the second example, the described image could retrieve similar images which include "woman" from image containing "woman eat", "lie" from image including "man lie" and also "couch" from another picture, thus helping generate a sentence with concept combinations of "woman lie". For the last one, we could find "horse

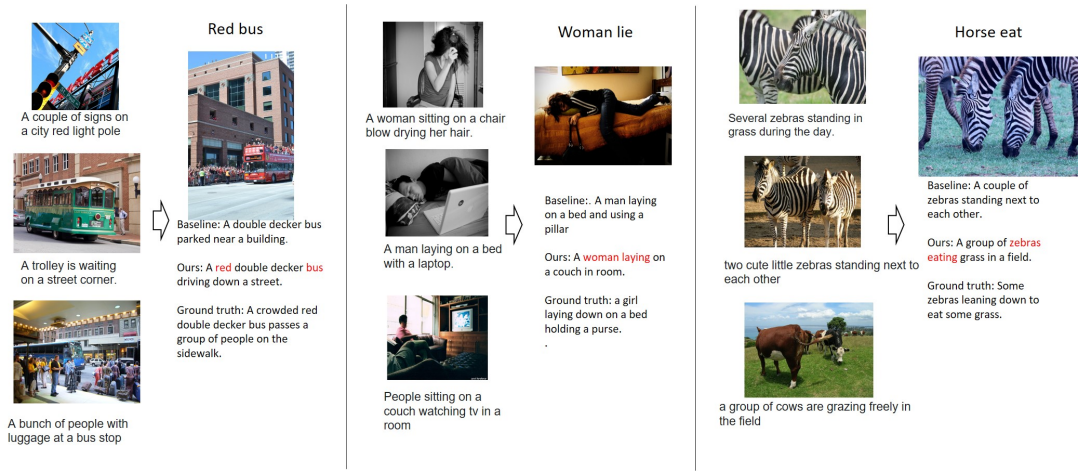

Figure 5: Examples of our prototype-based generation on the common concept split.

eat" combinations from "zebra stand" (zebra is categorized as synonym of "horse") and "cow grazing" (graze is categorized as synonym of "eat"), helping to generate "horse eat."

5 Discussion

5.1 How it helps different types of concept combinations

	Noun	Other	Combine	Total
Black cat	420	210	195	448
Big bird	94	56	40	123
Red bus	202	151	137	232
Small plane	149	25	21	158
Man eat	220	160	134	250
Woman lie	121	101	88	144

Table 6: Concept hit of the prototypes; Other means verb or attribute

Table 6 shows the hit rate of prototypes, e.g., for images containing "black cat" (448), the three prototypes can cover "cat" in 420 of the 448 images, "black" in 210 of 448, and both "black" and "cat" in 195 of 448 (note that "black" and "cat" are covered by different prototypes).

Attribute-Noun

(1) Color as the modifier: the attribute-noun pairs with color as the modifier have relatively good generalization performance, as shown in Table 4. Similar with other methods, we find that our model is better at generalizing to describe inanimate objects than animate objects as inanimate objects are more feature invariant.

(2) Size as the modifier: the generalization performance for size modifiers remains low for all models. It is because the size modifier has little correlation with the bounding box size; for example, a big bird could be very small in a image because it is viewed from a distance. It is more object or context dependent, e.g., a human has to grasp the commonsense knowledge of an average cat before describing a cat as small or large. Meanwhile, people sometimes need to reference other objects in the picture to describe the object of interest with a size modifier. In addition, we can also see from Table 6 that the retriever also fails to retrieve prototypes with the size modifier. Therefore it remains a hard question under this framework.

Noun-verb For these concept pairs, our method achieves a significant increase with regard to the baseline. Table 6 indicates that the hit rate of three prototypes covering the verbs is relatively higher than attributes. The increase of composition generalization could be attributed to the higher hit rate.

Rare concepts For the rare concepts, our method consistently improves the concept recall rate. It is due to the fact our retriever is capable to retrieve the concept from other training instances, thus upsampling that the rare concept. This can enhance the generation model with these rare concepts.

5.2 Why re-ranking helps

As illustrated from Table 2, re-ranking a large number of sentences produced by the beam search algorithm would significantly increase recall@5. We presume that the gain might be from a debiased decoding objective. The original objective is:

$$\hat{c} = \arg\max_{c} \log p(c \mid x) \tag{15}$$

To deduct the concept occurrence bias of captions in training so that the probability of sentences with

novel concepts would increase, we could therefore add a regularization term $\log p(c)$:

$$\hat{c} = \arg\max_{c}(\log p(c \mid x) - \lambda \log p(c)) \quad (16)$$

$$= \arg\max_{c}((1-\lambda)\log p(c \mid x) + \lambda \log p(x \mid c)) \quad (17)$$

However, directly decoding from Equation 17 is intractable as the second term $p(x|c)$ requires completion of caption generation before it can be computed. Practically, we turn to the re-ranking approach that involves first generating the top-n candidates based on the first term of the objective function and then re-ranking the top-n list using the other. As training a model to predict $p(x|c)$ is not trivial, empirically, we turn to the visual semantic similarity score $s(x, c)$ as an alternative [3].

6 Conclusion

We explore a prototype-based generation approach to encourage image captioning models to produce sentences with improved compositional generalization. We design a multi-prototype retriever and an analogical reasoning editor to merge the analogy entity pairs into the generation process. We demonstrate the effectiveness of the model on both composition related and quality related evaluation metrics over both common concept and rare concept splits. We perform detailed analyses on the results. In the future, we will explore this framework on the transformer based decoders.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957.

Ekin Akyürek, Afra Feyza Akyürek, and Jacob Andreas. 2020. Learning to recombine and resample data for compositional generalization. *arXiv preprint arXiv:2010.03706*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016a. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016b. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

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.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv*.

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.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907.

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*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improved visual-semantic embeddings. *arXiv*, 2(7):8.

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.

Dedre Gentner and Linsey Smith. 2012. Analogical reasoning. *Encyclopedia of human behavior*, 2:130–136.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

[3] $\exp(s(x, c))$ is proportion to $p_{retrieve}(x|c)$

Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. *arXiv preprint arXiv:1812.01194*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643.

Theo MV Janssen and Barbara H Partee. 1997. Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier.

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. *arXiv preprint arXiv:1912.09713*.

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. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

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 *ECCV*, pages 740–755. Springer.

Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-aware visual policy network for sequence-level image captioning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1416–1424.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 375–383.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *CVPR*, pages 7219–7228.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

James L McClelland, David E Rumelhart, PDP Research Group, et al. 1986. *Parallel distributed processing*, volume 2. MIT press Cambridge, MA.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. *arXiv preprint arXiv:1909.04402*.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1242–1250.

Yunchen Pu, Martin Renqiang Min, Zhe Gan, and Lawrence Carin. 2018. Adaptive feature abstraction for translating video to text. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2019. Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99.

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.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021a. Descriptive image captioning with salient retrieval priors. *Proceedings of the Canadian Conference on Artificial Intelligence*. Https://caiac.pubpub.org/pub/zllzroe5.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021b. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 269–277, Online. Association for Computational Linguistics.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7454–7464.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Stella Vosniadou and Andrew Ortony. 1989. *Similarity and analogical reasoning*. Cambridge University Press.

Bo Wu, Haoyu Qin, Alireza Zareian, Carl Vondrick, and Shih-Fu Chang. 2020. Analogical reasoning for visually grounded language acquisition. *arXiv preprint arXiv:2007.11668*.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694.

Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Russ R Salakhutdinov. 2016. Review networks for caption generation. In *Advances in neural information processing systems*, pages 2361–2369.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.