Object Counts! Bringing Explicit Detections Back into Image Captioning

Josiah Wang, Pranava Madhyastha and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{j.k.wang, p.madhyastha, l.specia}@sheffield.ac.uk

Abstract

The use of explicit object detectors as an intermediate step to image captioning – which used to constitute an essential stage in early work – is often bypassed in the currently dominant end-to-end approaches, where the language model is conditioned directly on a mid-level image embedding. We argue that explicit detections provide rich semantic information, and can thus be used as an interpretable representation to better understand why end-to-end image captioning systems work well. We provide an in-depth analysis of end-to-end image captioning by exploring a variety of cues that can be derived from such object detections. Our study reveals that end-to-end image captioning systems rely on matching image representations to generate captions, and that encoding the frequency, size and position of objects are complementary and all play a role in forming a good image representation. It also reveals that different object categories contribute in different ways towards image captioning.

1 Introduction

Image captioning (IC), or image description generation, is the task of automatically generating a sentential textual description for a given image. Early work on IC tackled the task by first running object detectors on the image and then using the resulting explicit detections as input to generate a novel textual description, e.g. (Kulkarni et al., 2011; Yang et al., 2011). With the advent of sequence-to-sequence approaches to IC, e.g. (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), coupled with the availability of large image description datasets, the performance of IC systems showed marked improvement, at least according to automatic evaluation metrics like Meteor (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015).

The currently dominant neural-based IC systems are often trained end-to-end, using parallel (image, caption) datasets. Such systems are essentially sequential language models conditioned directly on some mid-level image features, such as an image embedding extracted from a pre-trained Convolutional Neural Network (CNN). Thus, they bypass the explicit detection phase of previous methods and instead generate captions *directly* from image features. Despite significant progress, it remains unclear why such systems work. A major problem with these IC systems is that they are less interpretable than conventional pipelined methods which use explicit detections.

We believe that it is timely to again start exploring the use of explicit object detections for image captioning. Explicit detections offer rich semantic information, which can be used to model the entities in the image as well as their interactions, and can be used to better understand image captioning.

Recent work (Yin and Ordonez, 2017) showed that conditioning an end-to-end IC model on visual representations that implicitly encode object details yields reasonably good captions. Nevertheless, it is still unclear *why* this works, and *what* aspects of the representation allow for such a good performance. In this paper, we study end-to-end IC in the context of explicit detections (Figure 1) by exploring a variety of cues that can be derived from such detections to determine *what* information from such representations helps image captioning, and *why*. To our best knowledge, our work is the first experimental analysis of end-to-end IC frameworks that uses object-level information that is highly interpretable as a tool for understanding such systems. Our main contributions are as follows:

1. We provide an in-depth analysis of the performance of end-to-end IC using a simple, yet effective 'bag of objects' representation that
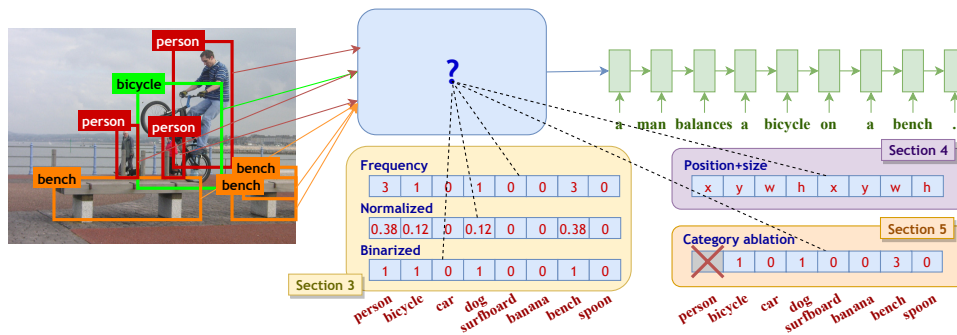
Figure 1: Using explicit detections as an intermediate step towards end-to-end image captioning. The question we investigate is what information can we extract from explicit detections that can be useful for image captioning.

is interpretable, and generates good captions despite being low-dimensional and highly sparse (Section 3).

2. We investigate whether other spatial cues can be used to provide information complementary to frequency counts (Section 3).

3. We study the effect of incorporating different spatial information of individual object instances from explicit detections (Section 4).

4. We analyze the contribution of the categories in representations for IC by ablating individual categories from them (Section 5).

Our hypothesis is that there are important components derived from explicit detections that can be used to effectively inform IC. Our study confirms our hypothesis, and that features such as the frequency, size and position of objects all play a role in forming a good image representation to match their corresponding representations in the training set. Our findings also show that different categories contribute differently to IC, and this partly depends on how likely they are to be mentioned in the caption given that they are depicted in the image. The results of our investigation will help further work towards more interpretable image captioning.

2 Related work

Early work on IC apply object detectors explicitly on an image as a first step to identify entities present in the image, and then use these detected objects as input to an image caption generator. The caption generator typically first performs content selection (selecting a subset of objects to be described) and generates an intermediate representation (e.g. semantic tuples or abstract trees), and then performs surface realization using rules, templates, n-grams or a maximum entropy language model. The main body of work uses object detectors for 20 pre-specified PASCAL VOC (Visual Object Classes) (Everingham et al., 2015) (Yang et al., 2011; Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012), builds a detector inferred from captions (Fang et al., 2015), or assumes gold standard annotations are available (Elliott and Keller, 2013; Yatskar et al., 2014).

Currently, deep learning end-to-end approaches dominate IC work (Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015). Such approaches do not use an explicit detection step, but instead use a 'global' image embedding as input (generally a CNN) and learn a language model (generally an LSTM) conditioned on this input. Thus, they are trained to learn image caption generation directly from a parallel image–caption dataset. The advantage is that no firm decisions need to be made about object categories. However, such approaches are hard to interpret and are dataset dependent (Vinyals et al., 2017).

Some recent work use object-level semantics for end-to-end IC (Gan et al., 2017; Wu et al., 2016; You et al., 2016). Such systems represent images as predictions of semantic concepts occurring in the image. These predictions, however, are at a global, *image* level ("does this image contain a chair?"), rather than at object *instance* level ("there is a big chair at position x"). In addition, most previous work regard surface-level terms extracted directly from captions as 'objects', while we use off-the-shelf predefined object categories which have a looser connection between the image and the caption (e.g. objects can be described in captions using different terms, depicted objects might not be mentioned in captions, and captions might mention objects that are not depicted).

Yin and Ordonez (2017) propose conditioning an end-to-end IC model on information derived from explicit detections. They implicitly encode the category label, position and size of object instances as an 'object-layout' LSTM and condition the language model on the final hidden state of this LSTM, and produce reasonably good image captions based only on those cues, without the direct use of images. Our work is different in that we feed information from explicit object detections *directly* to the language model in contrast to an object-layout LSTM which abstracts away such information, thereby retaining the *interpretability* of the input image representation. This gives us more control over the image representation which is simply encoded as a bag of categorical variables.

There is also recent work applying attention-based models (Xu et al., 2015) on explicit object proposals (Anderson et al., 2018; Li et al., 2017), which may capture object-level information from the attention mechanism. However, attention-based models require object information in the form of vectors, whereas our models use information of objects as categorical variables which allow for easy manipulation but are not compatible with the standard attention-based models. The model that we use, *under similar conditions* (i.e. under similar parametric settings), is comparable to the state-of-the-art models.

3 Bag of objects

We base our experiments on the MS COCO dataset (Lin et al., 2014). From our preliminary experiments, we found that a simple *bag of object categories* used as an image representation for end-to-end IC led to good scores according to automatic metrics, comparable to and perhaps even higher than those using CNN embeddings. This is surprising given that this bag of objects vector is *low-dimensional* (each element represents the frequency of one of 80 COCO categories) and *sparse* (mainly zeros, as only a few object categories tend to occur in a given image). In simple terms, it appears that the IC model can generate a reasonable caption by merely knowing what is in the image, e.g. that there are three *person*s, three *benches* and a *bicycle* in Figure 1.

This observation raises the following questions. What is it in this simple bag of objects representation that contributes to the surprisingly high performance on IC? Does it lie in the frequency counts? Or the choice of categories themselves? It is also worth noting that the image captions in COCO were crowd-sourced *independent* of the COCO object annotations, i.e. image captions were written based only on the image, without object-level annotations. The words used in the captions thus do not correspond directly to the 80 COCO categories (e.g. a *cup* may not be mentioned in a description even though it is present in the image, and vice versa, i.e. objects described in the caption may not correspond to any of the categories).

In order to shed some light into what makes bag of object categories representations work so well for IC, we first investigate whether the *frequency counts* is the main contributor. We then proceed to studying what else can be exploited from explicit object detections to improve on the bag of objects model, for example the size of object instances. We also perform an analysis on these representations to gain more insights into why the bag of objects model performs well.

3.1 Image captioning model

Our implementation is based on the end-to-end approach of Karpathy and Fei-Fei (2015). We use an LSTM (Hochreiter and Schmidhuber, 1997) language model as described in Zaremba et al. (2014). To condition the image information, we first perform a linear projection of the image representation followed by a non-linearity:

$$x = \sigma(W \cdot I_m) \quad (1)$$

where $I_m \in \mathcal{R}^d$ is the d-dimensional initial image representation, $W \in \mathcal{R}^{n \times d}$ is the linear transformation matrix, σ is the non-linearity. We use Exponential Linear Units (Clevert et al., 2016) as the non-linear activation in all our experiments. We initialize the LSTM-based caption generator with the projected image representation, x.

Training and inference. The caption generator is trained to generate sentences conditioned on x. We train the model by minimizing the cross-entropy, i.e. the sentence-level loss corresponds to the sum of the negative log likelihood of the correct word at each time step:

$$\Pr(S|x;\theta) = \sum_t \log(\Pr(w_t|w_{t-1}..w_0;x)) \quad (2)$$

where $\Pr(S|x;\theta)$ is the sentence-level loss con-

ditioned on the image feature x and $\Pr(w_t)$ is the probability of the word at time step t. This is trained with standard teacher forcing as described in Sutskever et al. (2014), where the correct word information is fed to the next state in the LSTM.

Inference is usually performed using approximate techniques like beam search and sampling methods. As we are mainly interested in studying different image representations, we focus on the language output that the models can most confidently produce. In order to isolate any other variables from the experiments, we generate captions using a greedy $\arg\max$ approach. We use a 2-layer LSTM with 128-dimensional word embeddings and 256-dimensional hidden dimensions. As training vocabulary we retain only words that appear at least twice. We provide details about hyperparameters and tuning in Appendix A.

3.2 Visual representations

The first part of our experiments studies the role of *frequency counts* of the 80-dimensional bag of objects representation. We explore the effects of using the following variants of the bag of objects representation: (i) **Frequency:** The number of instances per category; (ii) **Normalized:** The frequency counts normalized such that the vector sums to 1. This represents the *proportion* of object occurrences in the image; (iii) **Binarized:** An object category's entry is set to 1 if at least one instance of the category occurs, and 0 otherwise.

Berg et al. (2012) explore various factors that dictate what objects are mentioned in image descriptions, and found that object size and its position relative to the image centre are important. Inspired by these findings, we explore alternative representations based on these cues: (i) **Object size:** The area of the region provided by COCO, normalized by image size; we encode the largest object if multiple objects occur for the same category (max pooling). (ii) **Object distance:** The Euclidean distance from the object bounding box centre to the image centre, normalized by image size; we encode the object closest to the centre if multiple instances occur (min pooling). We also explore **concatenating** these features to study their complementarity.

Finally, we study the effects of **removing information** from the bag of objects representation. More specifically, we compare the results of retaining only a certain number of object instances

Representation	GT	Detect
CNN (ResNet-152 POOL5)	-	0.749
Frequency	0.807	0.752
Normalized	0.762	0.703
Binarized	0.751	0.703
Object min distance	0.759	0.691
Object max size	0.793	0.725
Obj max size + Obj min distance	0.799	0.743
Frequency + Obj min distance	0.830	0.769
Frequency + Obj max size	0.836	0.769
All three features	0.849	0.743

Table 1: CIDEr scores for image captioning using bag of objects variants as visual representations. We compare the results of using ground truth annotations (**GT**) and the output of a detector (**Detect**). As comparison we also provide, in the first row, the results of using a ResNet-152 POOL5 CNN image embedding with our implementation of an end-to-end IC system.

in the *frequency*-based bag of objects representation, rather than representing an image with all objects present. We experiment with retaining only the frequency counts for one object category and 25%, 50%, and 75% of object categories; the remaining entries in the vector are set to zero. The object categories to be retained are selected, per image: (i) randomly; (ii) by the $N\%$ most frequent categories of the image; (iii) by the $N\%$ largest categories of the image; (iv) by the $N\%$ categories closest to the centre of the image.

We performed these evaluations based on (i) ground truth COCO annotations and (ii) the output of an off-the-shelf object detector (Redmon and Farhadi, 2017) trained on 80 COCO categories. With ground truth annotations we can isolate issues stemming from incorrect detections.

3.3 Experiments

We train our models on the full COCO training set, and use the standard, publicly available splits[1] of the validation set as in previous work (Karpathy and Fei-Fei, 2015) for validation and testing (5,000 images each). We use CIDEr (Vedantam et al., 2015) – the official metric for COCO – as our evaluation metrics for all experiments. For completeness, we present scores for other common IC metrics in Appendix B.

Table 1 shows the CIDEr scores of IC systems using variants of the bag of objects representation, for both ground truth annotations and

---

[1] http://cs.stanford.edu/people/karpathy/deepimagesent

2183

Feature vs. Pooling	Min	Max	Mean
Obj. Size	0.748	0.793	0.789
Obj. Distance	0.759	0.768	0.740

Table 2: CIDEr scores for captioning comparing the use of min, max or average pooling of either object size or distance features, using ground truth annotations.

the output of an object detector. Compared to a pure CNN embedding (ResNet-152 POOL5), our object-based representations show higher (for ground truth annotations) or comparable CIDEr scores (for detectors). Our first observation is that frequency **counts** are essential to IC. Using normalized counts as a representation gives poorer results, which intuitively makes sense: An image with 20 cars and 10 people is significantly different from an image with two cars and one person. Using binarized counts (presence or absence) brings the score further down. This is to be expected: An image with one person is very different from one with 10 people.

Using spatial information (size or distance) also proved useful. Encoding the object size in place of frequency gave reasonably better results over using object distance from the image centre. We can conclude that the size and centrality of objects are important factors for captioning, with object size being more informative than position.

We also experimented with different methods for aggregating multiple instances of the same category, in addition to choosing the biggest instance and the instance closest to the image centre. For example, choosing the *smallest* instance (min pooling) or the instance *furthest* away from the image centre (max pooling), or just averaging them (mean pooling). Table 2 shows the results. For object size, the findings are as expected: Smaller object instances are less important for IC, although averaging them works comparably well. Surprisingly, in the case of distance, using the object *furthest* from the image centre actually gave slightly better results than the one closest. Further inspection revealed that aggregating instances is not effective in some cases. We found that the positional information (and interaction with other objects) captured by the object further away may sometimes represent the semantics of the image better than the object in the centre of the image. For example, in Figure 2, encoding only the position of the person in the middle will result in the

Obj. min distance:
● a man in a kitchen preparing food in a kitchen .
Obj. max distance:
● a group of people standing around a kitchen counter .

Figure 2: Example where encoding the distance of the object furthest away (solid green) is better than that of the one closest to the image centre (dashed red). The IC model assumes that only one person is in the middle in the former case, and infers that many people may be gathered around a table in the latter.

representation being similar to other images with only one person in the centre of the image (and also on a kitchen counter). Representing the person as the one furthest from the image will result in some inference (from training data) that there could be more than one person in the image sitting *around* the kitchen counter rather than a single person standing *at* the kitchen counter.

The combination of results (bottom row of Table 1) shows that the three features (frequency, object max size and min distance) are complementary, and that combining any pair gives better CIDEr scores than each alone. The combination of all three features produces the best results. These results are interesting, as adding spatial information of even just one object per category can produce a better score. This has, to our knowledge, not been previously demonstrated. The performance of using an explicit detector rather than ground truth annotations is poorer, as expected from noisy detections. However, the overall trend generally remains similar, except for the combination of all three features which gave poorer scores.

Finally, Figure 3 shows the results of partially removing or masking the information captured by the bag of object representation (frequency). As expected, IC performance degrades when less than 75% of information is retained. The performance of the system where the representation is reduced using frequency information suffers the most (even worst than removing categories randomly), suggesting that frequency does not correspond to an object category's importance, i.e. just because there is only one *person* in the image does
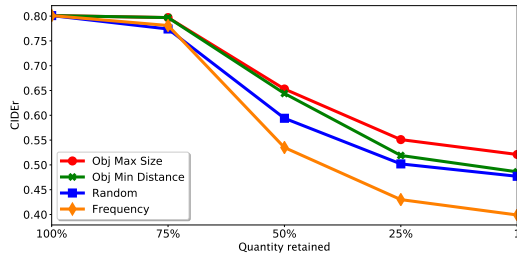
2184

Figure 3: Change in CIDEr scores for image captioning by reducing the number of (ground truth) object instances in the image representation, based on different heuristics.
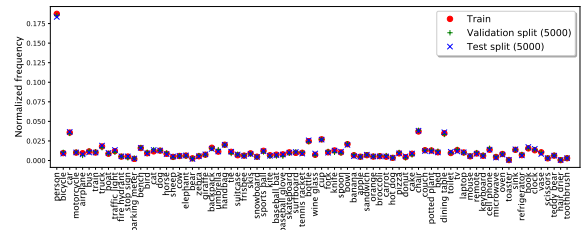

Figure 4: Object category distributions for COCO train, validation and test splits: normalized document frequency of each category. The distribution between the training and test sets are almost identical. A higher resolution version can be found in Appendix B.

not mean that it is less important than the ten *car*s depicted. On the other hand, object size correlates with object importance in IC, i.e. larger objects are more important than smaller objects for IC: The performance does not degrade as much as removing categories by their frequency in the image.

3.4 Analysis

We hypothesize that the bag of objects representation performs well because it serves as a good representation for the dataset and allows for better image matching. One observation is that the category distribution between the training and test sets are very similar (Figure 4), thus increasing the chance of the bag of objects representation producing a close match to one in the training set. From this observation, we posit that end-to-end IC models leverage COCO being repetitive to find similar matches for a test image to a combination of images in the training set. Further investigation on the category distribution (e.g. by splitting the dataset such that the test set contains unseen categories) is left for future work.

k-**Nearest neighbour analysis.** We further investigate our claim that end-to-end IC systems essentially perform complex image matching against the training set with the following experiment. The idea is that if the IC model performs some form of image matching and text retrieval from the training set, then the nearest neighbour (from training) of a test image should have a caption similar to the one generated by the model. However, the model does not always perform text retrieval as the LSTM is known to sometimes generate novel captions, possibly by aggregating or 'averaging' the captions of similar images and performing some factorization. We first generate captions for every training image using the bag of ob-

Type	BLEU	Meteor	CIDEr	SPICE
Freq.	0.868	0.591	6.956	0.737
Proj.	0.912	0.634	7.651	0.799
Exact (2301)	1.000	1.000	10.000	1.000
Freq. ($\neg$ Exact)	0.757	0.498	4.337	0.512
Proj. ($\neg$ Exact)	0.837	0.560	5.638	0.628

Table 3: k-Nearest Neighbour (k=5) trial on the ground truth bag of objects (**Freq.**) and the projected bag of objects (**Proj.**) representations. The references are captions of 5-nearest images in each space. *Exact* represents a subset of 2301 samples where all the 5 neighbours have 0 distance (replicas) and $\neg$ represents nearest neighbours that are not replicas of the test image.

jects model (with ground truth frequency counts). We then compute the k-nearest training images for each given test image using both the bag of objects representation and its projection (Eq. 1). Finally, we compute the similarity score between the generated caption of the test image against all k-nearest captions. The similarity score measures how well a generated caption matches its nearest neighbour's captions. We expect the score to be high if the IC system generates an image similar to something 'summarized' from the training set.

As reported in Table 3, overall the captions seem to closely match the captions of 5 nearest training images. Further analysis showed that 2301 out of 5000 captions had nearest images at a zero distance, i.e., the same exact representation was seen at least 5 times in training (note that CIDEr gives a score of 10 only if the test caption and *all* references are the same). We found that among the non-exact image matches, the projected image representation better captures candidates in the training set than bag of objects. Figure 5 shows the five nearest neighbours of an example non-exact match and their generated captions in the

test		*person (5), cup (8), spoon (1), bowl (8), carrot (10), chair (6), dining table (3)* ⇒ a group of people sitting around a table with food .
1		*person (4), cup (4), spoon (1), bowl (5), chair (6), dining table (4)* ⇒ a woman sitting at a table with a plate of food .
2		*person (9), bottle (1), cup (6), bowl (4), broccoli (2), chair (5), dining table (3)* ⇒ group of people sitting at a table eating food .
3		*person (11), cup (2), bowl (4), carrot (6), cake (1), chair (4), dining table (1)* ⇒ a group of people sitting around a table with a cake .
4		*cup (1), spoon (1), bowl (9), carrot (10), chair (3), potted plant (1), dining table (1), vase (1)* ⇒ a table with a variety of food on it .
5		*cup (4), bowl (7), carrot (6), dining table (1)* ⇒ a table with bowls of food and vegetables .

Figure 5: Five nearest neighbours from the training set in the projected space, for an example test image's (top row) original bag of objects representation that does not have an exact match in the training. For each image, we show the ground truth categories (and frequencies in parenthesis) and the generated caption. More examples can be found in Appendix D.

projection space. Note that the nearest neighbours are an approximation since we do not know the exact distance metric derived from the LSTM. We observe that the captions for unseen representations seem to be interpolated from multiple neighbouring points in the projection space, but further work is needed to analyze the hidden representations of the LSTM to understand the language model and to give firmer conclusions.

4 Spatial information on instances

Here we further explore the effect of incorporating spatial information of object detections for IC. More specifically, we enrich the representations by encoding positional and size information for more object instances, rather than restricting the encoding to only one instance per category which makes the representation less informative.

4.1 Spatial representation

We explore encoding **object instances and their spatial properties** as a fixed-size vector. In contrast to Section 3, we propose handling multiple instances of the same category by encoding spatial properties of *individual* instances rather than aggregating them as a single value. Each instance is represented as a tuple (x, y, w, h, a), where x and y are the coordinates of the centre of the bounding box and are normalized to the image width

Feature set	Fixed	Tuned
Bag of objects	0.807	0.834
(x, y, w, h, a)	0.870	0.915
(x, y, w, h)	0.859	0.898
(x, y, a)	0.850	0.900
(w, h)	0.870	0.920
(a)	0.869	0.857
(x, y)	0.810	0.863
LSTM Yin and Ordonez (2017)[†]		0.922

Table 4: CIDEr scores for image captioning using representations encoding spatial information of instances derived from ground truth annotations, with either fixed hyperparameters (Section 3.1) or with hyperparameter tuning. † Results taken from (Yin and Ordonez, 2017).

and height respectively, w and h are the width and height of the bounding box respectively, and a is the area covered by the object segment and normalized to the image size. Note that $w \times h \geq a$ (box encloses the segment). We assume that there are maximum 10 instances per vector, and instances of the same category are ordered by a (largest instance first). We encode each of the 80 categories as separate sets. Non-existent objects are represented with zeros. The dimension of the final vector is 4000 ($80 \times 10 \times 5$). We also perform a **feature ablation** experiment to isolate the contribution of different spatial components.

4.2 Experiments

All experiments in this subsection use ground truth annotations – we expect the results of using an object detector to be slightly worse but in most cases follow a similar trend, as shown in the previous section. Table 4 shows the CIDEr scores using the same setup as Section 3, but using representations with spatial information about individual object instances. Encoding spatial information led to substantially better performance over bag of objects alone. Consistent with our previous observation, w and h (bounding box width and height) seems to be the most informative feature combination – it performs well even without positional information. Area (a) is less informative than the combination of w and h, possibly because it compresses width-height ratio information despite discarding noise from background regions. Positional information (x, y) does not seem to be as informative, consistent with observations from previous work (Wang and Gaizauskas, 2016).

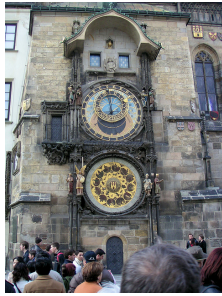The last column in Table 4 shows the CIDEr

Objects in the image	*person*, *clock*
Representation	**Caption**
Frequency	a large clock tower with a large clock on it .
Object min distance	a clock tower with a large clock on it 's face .
Object max size	a man standing in front of a clock tower .
All three features	a clock tower with people standing in the middle of the water .
(x, y)	a large clock tower with a clock on the front .
(w, h)	a clock on a pole in front of a building
(a)	a large clock tower with people walking around it
(x, y, w, h, a)	a group of people standing around a clock tower .
CNN (ResNet-152)	a large building with a clock tower in the middle of it .
person removed	a clock tower with a weather vane on top of it .

Image ID: 378657

Figure 6: Example captions with different models. The models with explicit object detection and additional spatial information ((x, y, w, h, a)) are more precise in most cases. The output of a standard ResNet-152 POOL5 is also shown, as well as that of the model where the most salient category – *person* – is removed from the feature vector. More example outputs are available in Appendix C.

scores when training the models by performing hyperparameter tuning during training. We note that the results with our simpler image representation are comparable to the ones reported in Yin and Ordonez (2017), which use more complex models to encode similar image information. Interestingly, we observe that positional information (x, y) work better than before tuning in this case.

Example outputs from the models in Sections 3 and 4 can be found in Figure 6.

5 Importance of different categories

In the previous sections, we explore IC based on explicit detections for 80 object categories. However, not all categories are made equal. Some categories could impact IC more than others (Berg et al., 2012). In this section we investigate **which categories are more important** for IC on the COCO dataset. Our **category ablation** experiment involves removing *one* category from the 80-dimensional bag of objects (ground truth frequency) representation at a time, resulting in 80 sets of 79D vectors without each ablated category.

We postulate that salient categories should lead to larger performance degradation than others. However, what makes a category 'salient' in general (*dog* vs. *cup*)? We hypothesize that it could be due to (i) how frequently it is depicted across images; (ii) how frequently it is mentioned in the captions when depicted in the image. To quantify these hypotheses, we **compute the rank correlation** between changes in CIDEr from removing the category and each of the statistic below:

- $f(v_c) = \sum_i^N 1(c \in C_i)$: frequency of the ablated category c being annotated among N images in the training set, where C_i is the

set of all categories annotated in image i, and $1(x)$ is the indicator function.

- $p(t_c|v_c) \approx \frac{f(t_c, v_c)}{f(v_c)}$: proportion of ablated category being mentioned in any of the reference captions given that it is annotated in the image in the training set.

For determining whether a depicted category is mentioned in the caption, the matching method described in Ramisa et al. (2015) is used to increase recall by matching category labels with (i) the term themselves; (ii) the head noun for multiword expressions; (iii) WordNet synonyms and hyponyms. We treat these statistics as an approximation because of the potential noise from the matching process, although it is clean enough for our purposes.

We have also tried computing the correlation with $f(t_c)$ (frequency of the category being mentioned regardless of whether or not it is depicted). However, we found the word matching process too noisy as it is not constrained or grounded on the image (e.g. "hot dog" is matched to the *dog* category). Thus, we do not report the results for this.

5.1 Experiments

Figure 7 shows the result of the category ablation experiment. Categories like *train*, *sandwich*, *person* and *spoon* led to the largest drop in CIDEr scores. On the other end, categories like *surfboard*, *carrot* and *book* can be removed without negatively affecting the overall score.

By comparing the CIDEr score changes against the frequency counts of object annotations in the training set (top row), there does not seem to be a clear correlation between depiction frequency and CIDEr. Categories like *bear* are infrequent but led to a large drop in score; likewise, *chair* and *dining table* are frequent but do not affect the results
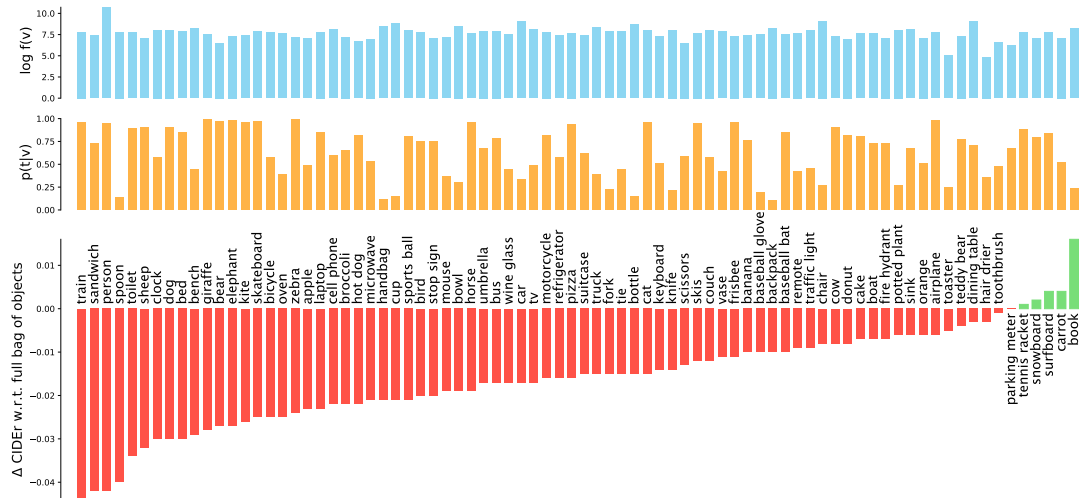
Figure 7: Difference in CIDEr scores when removing each category from the bag of objects representation (79 dimensions), compared to using the full 80D vector (bottom plot). See main text for details.

| | Coefficient value (p-value) | |
	$f(v_c)$	$p(t_c\|v_c)$
Spearman's ρ	0.137 (0.226)	0.227 (0.043)
Kendall's τ	0.093 (0.223)	0.153 (0.047)

Table 5: Correlation between changes in CIDEr score from category ablation and the frequency of depiction of the category ($f(v_c)$) against the probably of it being mentioned in the caption given depiction (($p(t_c|v_c)$).

as negatively. In contrast, the frequency of a category being mentioned given that it is depicted is a better predictor for the changes in CIDEr scores in general (middle row). Animate objects seem to be important to IC and are often mentioned in captions (Berg et al., 2012). Interestingly, removing *spoon* greatly affects the results even though it is not frequent in captions.

Table 5 presents the rank correlation (Spearman's ρ and Kendall's τ, two-tailed test) between changes in CIDEr and the two heuristics. While both heuristics are positively correlated with the changes in CIDEr, we can conclude that the frequency of being mentioned (given that it is depicted) is better correlated with the score changes than the frequency of depiction. Of course, the categories are not mutually exclusive and object co-occurrence may also play a role. However, we leave this analysis for future work.

Figure 6 shows an example when the category *person* is removed from the feature vector. Here, the model does not generate any text related to *per-

son, as the training set contains images of clocks without people in it.

6 Conclusions

In this paper we investigated end-to-end image captioning by using highly interpretable representations derived from explicit object detections. We provided an in-depth analysis on the efficacy of a variety of cues derived from object detections for IC. We found that frequency counts, object size and position are informative and complementary. We also found that some categories have a bigger impact on IC than others. Our analysis showed that end-to-end IC systems are image matching systems that project image representations into a learned space and allow the LSTM to generate captions for images in that projected space.

Future work includes (i) investigating how object category information can be better used or expanded to improve IC; (ii) analyzing end-to-end IC systems by using interpretable representations that rely on other explicit detectors (e.g. actions, scenes, attributes). The use of such explicit information about object instances could help improve our understanding of image captioning.

Acknowledgments

References

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 VQA. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*. IEEE.

Alexander C. Berg, Tamara L. Berg, Hal Daumé III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, and Kota Yamaguchi. 2012. Understanding and predicting importance in images. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3562–3569. IEEE.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (ELUs). In *Proceedings of the International Conference on Learning Representation (ICLR)*. arXiv.org.

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.

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 *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 2625–2634. IEEE.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1292–1302. Association for Computational Linguistics.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2015. The PASCAL Visual Object Classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1473–1482. IEEE.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1141–1150. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3128–3137. IEEE.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1601–1608. IEEE.

Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. 2017. Image caption with global-local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 4133–4139. AAAI.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 220–228. 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer International Publishing.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–220. Association for Computational Linguistics.

Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 6517–6525. IEEE.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 4566–4575. IEEE.

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 & Pattern Recognition (CVPR)*, pages 3156–3164. IEEE.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663.

Josiah Wang and Robert Gaizauskas. 2016. Don't mention the shoe! A learning to rank approach to content selection for image description generation. In *Proceedings of the Ninth International Natural Language Generation Conference (INLG)*, pages 193–202. Association for Computational Linguistics.

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 & Pattern Recognition (CVPR)*, pages 203–212. IEEE.

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 *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057. PMLR.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454. Association for Computational Linguistics.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 110–120. Association for Computational Linguistics and Dublin City University.

Xuwang Yin and Vicente Ordonez. 2017. Obj2Text: Generating visually descriptive language from object layouts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 177–187. Association for Computational Linguistics.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 4651–4659. IEEE.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR*, abs/1409.2329.

A Hyperparmeter Settings

The hyperparameter settings for our model are as follows:

- **LSTM layers**: 2-Layer LSTM
- **Word Embedding Dimensionality**: 128
- **Hidden Layer Dimensionality**: 256
- **Maximum Epochs**: 50
- **Batch Size**: [50, 100]
- **LSTM dropout settings**: [0.2, 0.7]
- **Vocabulary threshold**: 2
- **Learning Rate**: [1e-4, 4e-4]
- **Optimiser**: Adam

For items in a range of values, we used grid search to tune the hyperparmeters.

B Full Experimental Results

Tables 6 and 7 show the results of several of our experiments with the most common metrics used in image captioning: BLEU, Meteor, ROUGE$_L$, CIDEr and SPICE.

Figure 8 gives a high resolution version of Figure 4, showing the similarity between train and test distributions in terms of object categories.

C Example captions for different models

Figure 9 shows example images from COCO and the output captions from different models. We compare the outputs of selected models from Sections 3 and 4, and a model where the *person* category is removed from the input vector (Section 5).

D Example nearest neighbours for test images

Figure 10 shows the five nearest neighbours in the training set of each non-replica example from the test set (where the exact ground truth frequency representation does not occur in training). See Section 3.4 for a more detailed description of the experiment.

Representation	B1	B2	B3	B4	M	R	C	S
ResNet-152 POOL5	0.664	0.480	0.335	0.233	0.220	0.486	0.749	0.150
Frequency	0.668	0.481	0.334	0.231	0.223	0.486	0.807	0.155
Normalized	0.656	0.468	0.324	0.226	0.218	0.477	0.762	0.148
Binarized	0.652	0.465	0.317	0.217	0.218	0.473	0.751	0.146
Object min distance	0.650	0.460	0.316	0.219	0.218	0.474	0.759	0.147
Object max size	0.661	0.476	0.332	0.232	0.224	0.483	0.793	0.151
Obj max size + Obj min distance	0.670	0.482	0.333	0.231	0.225	0.485	0.799	0.153
Frequency + Obj min distance	0.675	0.491	0.345	0.239	0.229	0.495	0.836	0.160
Frequency + Obj max size	0.684	0.496	0.349	0.244	0.228	0.495	0.830	0.159
All three features	0.683	0.501	0.355	0.250	0.229	0.498	0.849	0.162

Table 6: Full results for image captioning using ground truth bag of objects variants as visual representations, for metrics **B**LEU, **M**eteor, **R**OUGE$_L$, **C**IDEr and **S**PICE.

Representation	B1	B2	B3	B4	M	R	C	S
Bag of objects	0.668	0.481	0.334	0.231	0.223	0.486	0.807	0.155
(x, y, w, h, a)	0.683	0.503	0.359	0.255	0.233	0.503	0.870	0.163
(x, y, w, h)	0.687	0.503	0.355	0.251	0.233	0.501	0.859	0.166
(x, y, a)	0.683	0.502	0.356	0.250	0.232	0.501	0.850	0.164
(w, h)	0.693	0.511	0.364	0.256	0.233	0.505	0.870	0.165
(a)	0.684	0.503	0.358	0.254	0.232	0.501	0.869	0.162
(x, y)	0.675	0.488	0.341	0.237	0.224	0.491	0.810	0.155

Table 7: Full results for image captioning, using representations encoding spatial information of instances derived from ground truth annotations with fixed hyperparameters, for metrics **B**LEU, **M**eteor, **R**OUGE$_L$, **C**IDEr and **S**PICE.
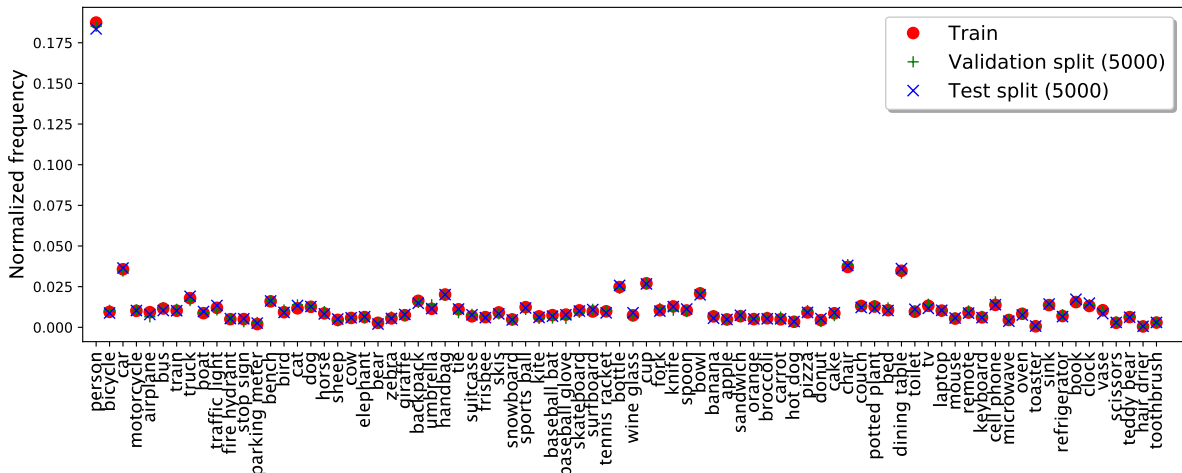
Figure 8: Object category distributions for COCO train, validation and test splits: normalized document frequency of each category. The distribution between the training and test sets are almost identical.
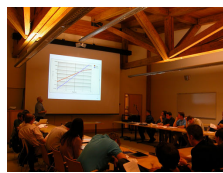
Image ID: 165225

Objects in the image	person, tie, tv, laptop, chair
Representation	**Caption**
Frequency	a group of people sitting around a table with a laptop .
Object min distance	a man sitting at a desk with a laptop computer .
Object max size	a man standing in front of a tv in a living room .
All three features	a group of people sitting at a table with a laptop .
(x, y)	a group of people standing around a table with a microphone .
(w, h)	a group of people sitting around a table with a laptop .
(a)	a group of people standing in a living room .
(x, y, w, h, a)	a group of people sitting at a table with laptops .
CNN (ResNet-152)	a group of people sitting around the table .
person removed	a man standing in front of a laptop computer .

Image ID: 196715

Objects in the image	person, car, truck, surfboard
Representation	**Caption**
Frequency	a man riding a surfboard on top of a wave .
Object min distance	a man riding a surfboard on top of a wave .
Object max size	a red and white truck driving down a street .
All three features	a red and white truck driving down a street .
(x, y)	a large white car parked in a parking lot .
(w, h)	a large white truck with a surfboard on the back of it .
(a)	a man riding a horse on a beach next to a dog .
(x, y, w, h, a)	a car is parked on the side of a street with a car in the background .
CNN (ResNet-152)	a boat is parked on the side of the road .
person removed	a man is standing on a surfboard in the water .

Image ID: 491793

Objects in the image	person, car, motorcycle, bus, handbag
Representation	**Caption**
Frequency	a group of people riding motorcycles down a street .
Object min distance	a city bus driving down a street next to a traffic light .
Object max size	a bus is driving down the street in the city .
All three features	a group of people standing around a parking lot .
(x, y)	a group of people standing around motorcycles in a parking lot .
(w, h)	a group of people standing around a bus stop .
(a)	a group of people riding bikes down a street .
(x, y, w, h, a)	a group of motorcycles parked next to each other on a street .
CNN (ResNet-152)	a bus is driving down a street with a lot of people .
person removed	a group of people riding motorcycles on a city street .

Image ID: 378657

Objects in the image	person, clock
Representation	**Caption**
Frequency	a large clock tower with a large clock on it .
Object min distance	a clock tower with a large clock on it 's face .
Object max size	a man standing in front of a clock tower .
All three features	a clock tower with people standing in the middle of the water .
(x, y)	a large clock tower with a clock on the front .
(w, h)	a clock on a pole in front of a building
(a)	a large clock tower with people walking around it
(x, y, w, h, a)	a group of people standing around a clock tower .
CNN (ResNet-152)	a large building with a clock tower in the middle of it .
person removed	a clock tower with a weather vane on top of it .

Figure 9: Examples of descriptions where models differ. The models with explicit object detection and additional spatial information $((x, y, w, h, a))$ is more precise in most cases (even though still incorrect in the second example). In the first example, aggregating multiple instances for size and distance cues clearly removes the information about the group of people in the image. The output of a standard CNN (ResNet-152 POOL5) is also shown, as well as that of the model where the most salient category – *person* – is removed.

Test Image ID: 242946

test		*person (5), cup (8), spoon (1), bowl (8), carrot (10), chair (6), dining table (3)* ⇒ a group of people sitting around a table with food .
1		*person (4), cup (4), spoon (1), bowl (5), chair (6), dining table (4)* ⇒ a woman sitting at a table with a plate of food .
2		*person (9), bottle (1), cup (6), bowl (4), broccoli (2), chair (5), dining table (3)* ⇒ a group of people sitting at a table eating food .
3		*person (11), cup (2), bowl (4), carrot (6), cake (1), chair (4), dining table (1)* ⇒ a group of people sitting around a table with a cake .
4		*cup (1), spoon (1), bowl (9), carrot (10), chair (3), potted plant (1), dining table (1), vase (1)* ⇒ a table with a variety of food on it .
5		*cup (4), bowl (7), carrot (6), dining table (1)* ⇒ a table with bowls of food and vegetables .

Test Image ID: 378962

test		*person (14), backpack (3), umbrella (4), handbag (1), banana (4), apple (6), orange (10), chair (7), dining table (2)* ⇒ a group of people standing around a fruit stand .
1		*person (14), backpack (1), banana (5), apple (5), orange (13)* ⇒ a group of people standing around a fruit stand .
2		*person (13), truck (1), backpack (1), apple (2), orange (10)* ⇒ a group of people standing around a fruit stand .
3		*person (12), bicycle (1), handbag (3), banana (2), apple (2), orange (7)* ⇒ a group of people standing around a fruit stand .
4		*person (11), banana (4), apple (1), orange (14)* ⇒ a man standing next to a fruit stand with bananas .
5		*person (14), backpack (1), handbag (3), apple (7), orange (14)* ⇒ a group of people standing around a fruit stand .

Test Image ID: 223648

test		*fork (3), spoon (14), chair (6), dining table (1), book (14)* ⇒ a dining room with a table and chairs .
1		*fork (1), knife (1), spoon (13), scissors (1)* ⇒ a drawer of a variety of different types of food .
2		*person (1), cup (3), knife (1), spoon (14), bowl (1), potted plant (1), dining table (1), vase (1)* ⇒ a woman is sitting at a table with a glass of wine .
3		*person (1), cup (1), fork (3), spoon (6), chair (9), dining table (5), vase (1)* ⇒ a woman sitting at a table with a plate of food .
4		*person (3), bottle (3), wine glass (5), cup (4), fork (3), knife (2), spoon (11), bowl (5), chair (4), dining table (2), book (2)* ⇒ a group of people sitting around a table with food .
5		*fork (4), knife (4), spoon (7), chair (1), couch (2), dining table (1)* ⇒ a table with a table and chairs and a table .

Figure 10: Five nearest neighbours from the training set in the projected space, for several example test images' (top row of each table) original bag of objects representation that does not have an exact match in the training. For each image, we show the ground truth categories (and frequencies in parenthesis) and the generated caption.