

# Déjà Image-Captions: A Corpus of Expressive Descriptions in Repetition

Jianfu Chen<sup>†</sup> and Polina Kuznetsova<sup>†</sup> and David S. Warren<sup>†</sup> and Yejin Choi<sup>‡</sup>  
Stony Brook University<sup>†</sup> University of Washington<sup>‡</sup>

{jianchen,pkuznetsova,warren}@cs.stonybrook.edu<sup>†</sup>, yejin@cs.washington.edu<sup>‡</sup>

## Abstract

We present a new approach to harvesting a large-scale, high quality image-caption corpus that makes a better use of already existing web data with no additional human efforts. The key idea is to focus on *Déjà Image-Captions: naturally* existing image descriptions that are *repeated almost verbatim* – by more than one individual for different images. The resulting corpus provides association structure between 4 million images with 180K unique captions, capturing a rich spectrum of everyday narratives including figurative and pragmatic language. Exploring the use of the new corpus, we also present new conceptual tasks of visually situated paraphrasing, creative image captioning, and creative visual paraphrasing.

## 1 Introduction

The use of multimodal web data has been a recurring theme in many recent studies integrating language and vision, e.g., image captioning (Ordonez et al., 2011; Hodosh et al., 2013; Mason and Charniak, 2014; Kuznetsova et al., 2014), text-based image retrieval (Rasiwasia et al., 2010; Rasiwasia et al., 2007), and entry-level categorization (Ordonez et al., 2013; Feng et al., 2015).

However, much research integrating complex textual descriptions to date has been based on datasets that rely on substantial human curation or annotation (Hodosh et al., 2013; Rashtchian et al., 2010; Lin et al., 2014), rather than using the web data in the wild as is (Ordonez et al., 2011; Kuznetsova et al., 2014). The need for human curation limits the potential scale of the multimodal dataset. Without human curation, however, the web data introduces significant noise. In particular, everyday captions

often contain extraneous information that is not directly relevant to what the image shows (Kuznetsova et al., 2013b; Hodosh et al., 2013).

In this paper, we present a new approach to harvesting a large-scale, high quality image-caption corpus that makes a better use of already existing web data with no additional human efforts. Figure 1 shows sample captions in the resulting corpus, e.g., “*butterfly resting on a flower*” and “*evening walk along the beach*”. Notably, some of these are figurative, e.g., “*rippled sky*” and “*sun is going to bed.*”

The key idea is to focus on *Déjà Image-Captions*, i.e., naturally existing image captions that are *repeated almost verbatim* by more than one individual for different images. The hypothesis is that such captions represent common visual content across multiple images, hence are more likely to be free of unwanted extraneous information (e.g., specific names, time, or any other personal information) and better represent visual concepts. A surprising aspect of our study is that such a strict data filtration scheme can still result in a large-scale corpus; sifting through 760 million image-caption pairs, we harvest as many as 4 million image-caption pairs with 180K unique captions.

The resulting corpus, *Déjà Image Captions*, provides several unique properties that complement human-curated or crowd-sourced datasets. First, as our approach is fully automated, it can be readily applied to harvesting a new dataset from the ever changing multimodal web data. Indeed, a recent internet report estimates that billions of new photographs are being uploaded daily (Meeker, 2014). In contrast, human-annotated datasets are costly to scale to different domains.

Second, datasets that are harvested from the web



Figure 1: The image-caption association graph of *Déjà Image-Captions*. Solid lines represent original captions and dotted lines represent paraphrase captions. This corpus reflects a rich spectrum of everyday narratives people use in online activities including figurative language (e.g., “*Sun is going to bed*”), casual language (e.g., “*Chillaxing at the beach*”), and conversational language (e.g., “*Can you spot the butterfly*”). The numbers in the parenthesis show the cardinality of images associated with each caption. Surprisingly, some of these descriptions are highly expressive, almost *creative*, and yet not unique — as all these captions are repeated almost verbatim by different individuals describing different images.

can complement those based on prompted human annotations. The latter in general are literal and mechanical readings of the visual scenes, while the former reflect a rich spectrum of natural language utterances in everyday narratives, including figurative, pragmatic, and conversational language, e.g., “*can you spot the butterfly*” (Figure 1). Therefore, this dataset offers unique opportunities for grounding figurative and metaphoric expressions using visual context.

In conjunction with the new corpus, publicly shared at <http://www.cs.stonybrook.edu/~jianchen/deja.html>, we also present three new tasks: *visually situated paraphrases* (§5); *creative image captioning* (§7), and *creative visual paraphrasing* (§7). The central algorithm component in addressing all these tasks is a simple and yet effective approach to image caption transfer that exploits the unique association structure of the resulting corpus (§3).

Our empirical results collectively demonstrate that when the web data is available at such scale, it is possible to obtain a large-scale, high-quality dataset with significantly less noise. We hope that our approach would be only one of the first attempts, and inspire future research to develop better ways of making use of ever-growing multimodal web data. Although it is unlikely that the automatically gathered datasets can completely replace the curated descriptions written in a controlled setting, our hope is to find ways to complement human annotated datasets in terms of both the scale and also the diversity of the domain and language.

The remainder of this paper is organized as follows. First we describe the dataset collection procedure and insights (§2). We then present a new approach to image caption transfer based on the association structure of the corpus (§3) followed by experimental results (§4). After then we present new conceptual tasks: *visual paraphrasing* (§5), *creative image captioning*, and *creative visual paraphrasing* (§7), interleaved with corresponding experimental results (§6, §8).

## 2 Dataset - Captions in Repetition

Our corpus consists of three components (Table 1): **MAIN SET** The first step is to crawl as many image-caption pairs as possible. We use flickr.com search API to crawl 760 million pairs in total. The API allows searching images within a given time window, which enables exhaustive search over any time span. To ensure visual correspondence between images and captions, we set query terms using 693 most frequent nouns from the dataset of Ordonez et al. (2011), and systematically slide time windows over the year 2013.<sup>1</sup> For each image, we segment its title and the first line of its description into sentences.

The crawled dataset at this point includes a lot of noise in the captions. Hence we apply initial filtering rules to reduce the noise. We retain only those image-sentence pairs in which the sentence contains the query noun, and does not contain personal information indicators such as first-person pronouns. We

<sup>1</sup>To ensure enough number of images are associated with each caption, we further search captions with no more than 10 associated images across *all* years.

set	# captions	# images
MAIN	176,780	3,967,524
PARAPHRASE	7,570 human-annotated triples 353,560 auto-generated triples	
FIGURATIVE	6,088 quotations 18,179 quotations + predicted figurative captions	180,185 413,698

Table 1: Corpus Statistics

	mean	std	25%	50%	75%	max
#imgs.	22.4	47.6	4	10	25	4617
#tokens	4.9	3.3	3	4	5	178

Table 2: Percentiles of the image count associated with each caption and the number of tokens in each caption.

want captions that are more than simple keywords, thus we discard trivial captions that do not include at least one verb, preposition, or adjective.

The next step is to find captions in repetition. For this purpose, we transform captions into *canonical forms*. We lemmatize all words, convert prepositions to a special token “IN”<sup>2</sup>, and discard function words, numbers, and punctuations. For instance, “*The bird flies in blue sky*” and “*A bird flying into the blue sky*” have the same canonical form, “*bird fly IN blue sky*”. We then retain only those captions that are repeated with respect to their canonical forms by more than one user, and for distinctly different images to ensure the generality of the captions.

Retaining only captions that are repeated verbatim may seem overly restrictive. Nonetheless, because we start with as many as 760 million pairs, this procedure yields nearly 180K unique captions associated with nearly 4M images.<sup>3</sup> What is more surprising, as will be shown later, is that many of these captions are highly expressive. Table 2 shows the distribution of the number of images associated with each caption.<sup>4</sup> The median and mean are 10 and 22.4 respectively, showing a high degree of connectivities between captions and images.

**PARAPHRASE SET** Our dataset collection procedure finds *one-to-many* relations between captions

<sup>2</sup>We do this transformation so as not to over-count unique captions with trivial variations, but merging prepositions can sometimes combine prepositions that are not semantically compatible. We therefore also keep original captions with original prepositions.

<sup>3</sup>We also keep user annotated image tags if available.

<sup>4</sup>Without counting additional edges created by visual paraphrasing (§5).

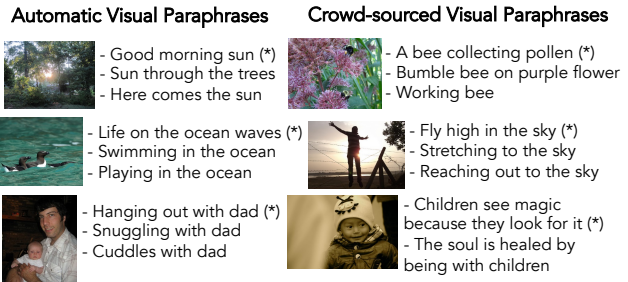


Figure 2: Example visual paraphrases: automatic (left) and crowd-sourced (right). The first caption marked with \* indicates the original caption of the corresponding image. Some paraphrases are not strictly equivalent to the original caption if considered out of context, while they are pragmatically adequate paraphrases given the image.

figure of speech	#caps.	example (#imgs.)
quotation&idiom	70	The early bird gets the worm (77)
personification	43	Meditating cat (38)
metaphor	24	Wine is the answer (7)
question	18	Do you see the moon (82)
dialog	11	Hello little flower (37)
anaphora	6	Beads, beads and more beads (62)
simile	5	The lake is like glass (23)
hyperbole	1	In the land of a billion lights (3)

Table 3: Distribution of figurative language out of 1000 random captions (171 figurative captions in total)

and images. To extend these relations to *many-to-many*, we introduce *visually-situated paraphrases* (or *visual paraphrases* for shorthand) (§5). A visual paraphrase relation is a triple  $(i, c, p)$ , where image  $i$  has an original caption  $c$ , caption  $p$  is the visual paraphrase for  $c$  situated in image  $i$ . We collect visual paraphrases for sample images in our dataset, using both crowd sourcing (7,570 triples) and an automatic algorithm (353,560 triples) (see §5 for details). Figure 2 shows example visual paraphrases.

Formally, our corpus represents a bipartite graph  $G = (T, V, E)$ , in which the set of captions  $T$  and the set of images  $V$  are connected by typed edges  $e(c, i, t)$ , where caption  $c \in T$ , image  $i \in V$ , and edge type  $t \in \{original, paraphrase\}$ , which denotes whether the image-caption association is given by the original caption or by a visual paraphrase.

**FIGURATIVE SET** We find that many repeating captions are surprisingly lengthy and expressive, most of which turn out to be idiomatic expressions and quotations, e.g., “*faith is the bird that feels the light when the dawn is still dark*” from Tagore’s poem. We look up goodreads.com

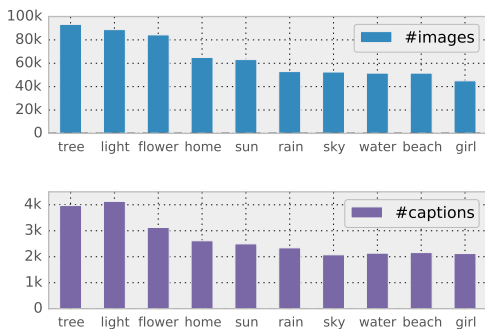


Figure 3: Top 10 queries with the largest number of images and unique captions

polarity	% in all caps.	mean/median #imgs. per cap.	example (#imgs)
pos.	8%	20 / 8	Happy bride and groom (282) The rock and pool, is nice and cool (4)
neg.	2%	19.5 / 7	Bad day at the office (269) Crying lightning (147)

Table 4: Distribution of caption sentiment. The polarity is determined by comparing number of positive words and negative words ( $>$ : positive;  $<$ : negative) according to a sentiment lexicon (Wilson et al., 2005) (counting only words of *strong* polarity).

and [brainyquotes.com](http://brainyquotes.com) to identify 6K quotation captions illustrated by 180K images. We also present a manual labeling on a small subset of the data (Table 3) to provide better insights into the degree and types of figurative speech used in natural captions. Using these labels we build a classifier (§7) to further detect 18K figurative captions associated with 410K images.

**INSIGHTS** As additional insights into the dataset, Figure 3 shows statistics of the visual content, Table 5 shows syntactic types of the captions, and Table 4 shows positive and negative sentiment in captions.

### 3 Image Captioning using Association Structure

We demonstrate the usefulness of the association between images and captions via retrieval-based image captioning. Given a query image  $q$  and the corpus  $G = (T, V, E)$ , the task is to find a caption  $c \in T$  that maximizes an affinity function  $\mathcal{A}(q, c)$ , which measures how well the caption  $c$  fits the query image  $q$ ,

$$c^* = \arg \max_{c \in T} \{\mathcal{A}(q, c)\} \quad (1)$$

**Visual Neighborhood:** Each textual description, e.g., “reading a book”, can associate with many dif-

type	%caps.	%imgs.	mean #imgs.	std #imgs.
verb	45%	44%	22	9
	be, have, do, look,		Sky is the limit (3057)	
	go, make, come, get,		Home is where the heart is (2480)	
	wait, take, love, play,		Lunch is served (2443)	
	walk, fly, see, watch,		Let them eat cake (2193)	
find, live, sleep, fall		Follow the yellow brick road (2077)		
prep	44%	41%	21	9
	in, of, on,		On the road (4617)	
	at, with, for,		After the rains (4450)	
	from, by,		Under the bridge (3443)	
	over, through		At the beach (3203)	
adj	11%	15%	30	15
	old, little, new,		Home sweet home (2398)	
	red, blue, more,		Good morning sun (1122)	
	white, big, beautiful,		Cabbage white butterfly (976)	
	black		Next door neighbors (838)	

Table 5: Statistics on the syntactic composition of captions. *verb*: captions with at least one verb. *prep*: prepositional phrases (without any verbs). *adj*: adjective phrases (without any verbs and prepositions). For each caption type, we also show the *top words* that appear in the most number of captions (left), and the *top captions* that are associated with largest number of images (right).

ferent visual instantiations (Figure 4a). Our dataset  $G = (T, V, E)$  serves as a database to navigate the possible visual instantiations of descriptive captions as observed in online photo sharing communities. Let  $\mathcal{N}_c = \{i | e(c, i, original) \in E\}$  denote the set of adjacent nodes (i.e., visual instantiations) of a caption  $c$ . To quantify how well a caption  $c$  describe a query image  $q$ , we propose to examine caption  $c$ ’s visual neighborhood  $\mathcal{N}_c$  as provided in our dataset. Concretely, the affinity  $\mathcal{A}(q, c)$  of a query image  $q$  to a caption  $c$  is a function  $\phi(q, \mathcal{N}_c)$  of  $q$  and the visual neighborhood  $\mathcal{N}_c$  defined as:

$$\mathcal{A}(q, c) = \phi(q, \mathcal{N}_c) = \frac{1}{\sigma} \sum_{i=1}^{\sigma} sim(q, \mathcal{N}_c^i) \quad (2)$$

where  $\sigma$  is a parameter;  $sim(\cdot, \cdot)$  is a similarity function of two images; and  $\mathcal{N}_c = [\mathcal{N}_c^1, \mathcal{N}_c^2, \dots, \mathcal{N}_c^{|\mathcal{N}_c|}]$  is sorted by  $sim(q, \mathcal{N}_c^i)$  in *descending* order.

Figure 4a illustrates the key insight: instead of directly transferring the caption of the single image with the closest visual similarity to the query image (Ordonez et al., 2011), we propose to retrieve a caption based on the aggregated visual similarity between its visual neighborhood and the query image. The idea is to prefer a caption for which the query image is likely to be a *prototypical* visual rendering (Ordonez et al., 2013; Deselaers and Ferrari, 2011), hence avoid an unusual association between the text

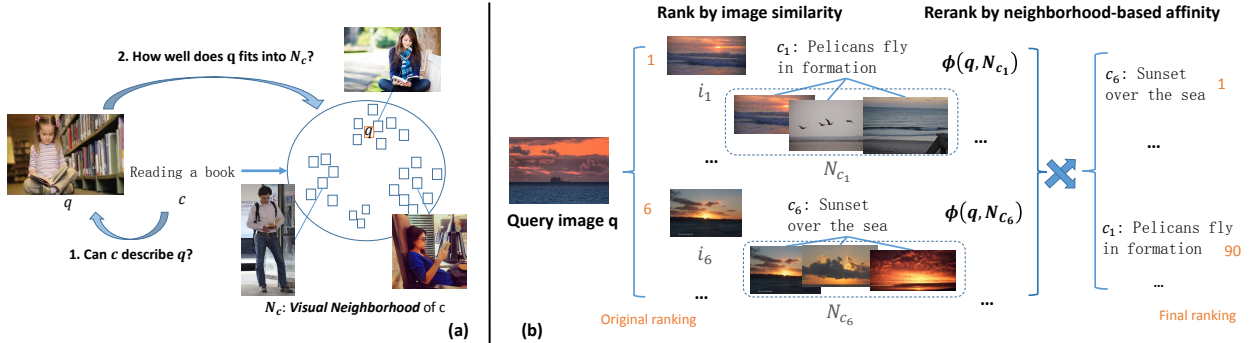


Figure 4: (a) Using the association structure, we retrieve a caption for which the query image is likely to be a *prototypical* visual rendering. We hypothesize that there can be multiple visual prototypes of a caption. (b) Reranking by visual neighborhood proximity.

and the visual information. Also, we hypothesize that there could be several diverse visual prototypes of any given textual description  $c$ , so we focus on only the top  $\sigma$  nearest members of  $\mathcal{N}_c$ .

We apply the neighborhood-based affinity for image captioning via reranking (Figure 4b): first we retrieve a pool of  $K$  candidate captions by finding top  $K$  closest images based on their direct visual similarity to the query image, then compute the neighborhood-based affinity to rerank the captions.<sup>5</sup> The proposed approach is similar in spirit to the non-parametric K nearest neighbor approach of (Boiman et al., 2008) in modeling image-to-concept similarity rather than image-to-image similarity, but differs in that our work is in the context of image description generation rather than classification.

#### 4 Experiments: Association Structure Improves Image Captioning

**Baselines:** The proposed approach (to be referred as ASSOC) requires one-to-many mappings between captions and images *at scale* — a unique property of our dataset. We compare against two baselines: instance-based retrieval of (Ordonez et al., 2011) (INSTANCE) and Kernel Canonical Correlation Analysis (KCCA) (Hardoon et al., 2004; Hodosh et al., 2013). We implement KCCA with Hardoon’s code<sup>6</sup>. We use a linear kernel since non-linear kernels like RBF showed worse performance.

<sup>5</sup>We set  $K = 100$  and choose parameter  $\sigma$  using a held-out development set of 300 images. If there are less than  $\sigma$  available images, we use them all.

<sup>6</sup>[http://www.davidroihardoon.com/Professional/Code\\_files/kcca\\_package.tar.gz](http://www.davidroihardoon.com/Professional/Code_files/kcca_package.tar.gz)

method	BLEU	METEOR
INSTANCE	0.125	0.029
KCCA	0.118	0.024**
ASSOC <sup>gi</sup> w/ all	0.130	0.031
ASSOC <sup>g+t</sup> w/ all	0.133	0.030
ASSOC <sup>ti</sup> w/ all	0.126	0.029
ASSOC <sup>gi</sup> w/ $\sigma$	0.172**	0.033*
ASSOC <sup>g+t</sup> w/ $\sigma$	0.159**	0.033*
ASSOC <sup>ti</sup> w/ $\sigma$	<b>0.184**</b>	<b>0.034**</b>

Table 6: Automatic evaluation for image captioning: The superscripts denote the image feature for reranking; gi: GIST; ti: Tinyimage; g+:= gi + ti. We report the best setting (gt) for INSTANCE and KCCA. Results statistically significant compared to INSTANCE with two-tailed  $t$ -test are indicated with \* ( $p < 0.05$ ) and \*\* ( $p < 0.005$ ).

**Configurations:** For image features, we follow (Ordonez et al., 2011) to experiment with two global image descriptors and their combination: a) the GIST feature that represents the dominant spatial structure of a scene (Oliva and Torralba, 2001); b) the Tinyimage feature that represents the overall color of an image (Torralba et al., 2008); c) a combination of the two. We compute the similarity as  $sim(Q, I) = -\|Q - I\|^2$ . The INSTANCE and the KCCA approaches use the feature combination. The ASSOC approach also use the combination for preparing candidate captions, but can use different features for reranking.

**Dataset:** We randomly sample 1000 images with unique captions as test set. The rest of the corpus is the pool of caption retrieval after *discarding*: (1) the original caption  $c$  and all of its associated images, to avoid potential unfair advantage toward ASSOC and (2) the 10K captions used for training KCCA and all

reranking feature	INSTANCE	ASSOC
gi	42%	58%
g+t	50%	50%
ti	46%	54%

Table 7: Human evaluation for image captioning: the % of cases judged as visually more *relevant*, in pairwise comparisons. gi: GIST; ti: Tinyimage; g+t:= gi+ti.

of their associated images (about 280K).

**Evaluation.** Automatic evaluation remains to be a challenge (Elliott and Keller, 2014). We report both BLEU (Papineni et al., 2002) at 1 without brevity penalty, and METEOR (Banerjee and Lavie, 2005) with balanced precision and recall. Table 6 shows the results: the ASSOC approach (w/  $\sigma$ ) significantly outperforms the two baselines. The largest improvement over INSTANCE is 60% higher in BLEU, and 44% higher in METEOR, demonstrating the benefit of the innate association structure of our corpus. Using *all* visual neighborhood (ASSOC w/ all) does not yield as strong results as selective neighborhood (ASSOC w/  $\sigma$ ), confirming our hypothesis that each visual concept can have diverse visual renderings.

We also compute crowd-sourced evaluation on a subset (200 images) randomly sampled out of the test set. For each query image, we present two captions generated by two competing methods in a random order. Turkers choose the caption that is more relevant to the visual content of the given image. We aggregate the choices of three turkers by majority voting. As shown in Table 7, ASSOC shows overall improvement over baselines, where the difference is more pronounced when reranking is based on feature sets that differ from the one used during the candidate retrieval.

## 5 Image Captioning using Visual Paraphrases

We present an exploration of *visually situated paraphrase* (or *visual paraphrase* in short hand), and demonstrate their utility for image captioning. Formally, given our corpus  $G = (T, V, E)$ , a visual paraphrase relation is a triple  $(i, c, p)$ , where given an image  $i \in V$  and its original caption  $c \in T$  (i.e.,  $e(c, i, original) \in E$ ),  $p \in T$  is a visual paraphrase for  $c$  situated in a visual context given by the image  $i$  (i.e.,  $e(p, i, paraphrase) \in E$ ). We collect visual paraphrases using both human annotation and an automatic algorithm.

### (1) Visual Paraphrasing using Crowd-sourcing:

We use Amazon Mechanical Turk to annotate visual paraphrases for a subset of images in our corpus. Given each image with its original caption, we showed 10 randomly sampled candidate captions from our dataset that share at least one physical-object noun<sup>7</sup> with the original caption. Turkers choose all candidate captions that could also describe the given image. We collect 7,570  $(i, c, p)$  paraphrase triples in total.

### (2) Visual Paraphrasing using Associative Structure:

We also propose an algorithm for automatic visual paraphrasing by adapting the ASSOC algorithm for image captioning (§3) as follows: given an image-caption pair  $(i, c)$ , it first prepares a set of candidate captions that share the largest number of physical-object nouns with  $c$ , which are likely to be semantically close to  $c$ ; then we rerank the candidate captions using the same neighborhood-based affinity as described in §3.

We apply this algorithm to generate a large set of visual paraphrases. For each caption in our corpus, we randomly sample two of its associated images, and generate one visual paraphrase for each image-caption pair, which yields 353,560  $(i, c, p)$  triples. See Figure 2 for example paraphrases.

## 5.1 Image Captioning using Visual Paraphrasing

We propose to utilize automatically-generated visual paraphrases to improve the ASSOC approach (§3) for image captioning. One potential limitation of the ASSOC approach is that for some captions, the number of associated images might be too small for reliable estimations of the neighborhood based affinity. We hypothesize that for a caption with a small visual neighborhood, merging its neighborhood with those associated with its *visual paraphrases* will give a more reliable estimation of the affinity between a query image and that caption. Thus we modify the ASSOC approach as follows.

After preparing a pool of  $K$  candidate captions  $\{c_1, c_2, \dots, c_K\}$ , automatically generate a visual paraphrase  $(i_i, c_i, p_i)$  for each  $(i_i, c_i)$ ; then rerank the candidate captions by the following affinity function that merges the visual neighborhood from the

<sup>7</sup>under the WordNet “physical\_entity.n.01” synset

method	BLEU	METEOR	AMT
INSTANCE	0.125	0.029	N/A
ASSOC <sup>gi</sup>	0.172	0.033	45%
ASSOC <sup>gi</sup> <sub>para</sub>	<b>0.187</b>	<b>0.036</b>	55%
ASSOC <sup>ti</sup>	0.184	0.034	45%
ASSOC <sup>ti</sup> <sub>para</sub>	<b>0.197</b>	<b>0.036</b>	55%

Table 8: Automatic and human evaluation of exploiting visual paraphrases for image captioning. The superscripts represent the image feature used in the reranking step; gi: GIST; ti: Tinyimage. The AMT column shows the percentages of captions preferred by human as of better visual relevance, in pairwise comparisons. The improvement of ASSOC<sub>para</sub> over ASSOC is significant at  $p < 0.002$  for BLEU, and  $p < 0.03$  for METEOR with two tailed  $t$ -test.

paraphrase,

$$\mathcal{A}(q, C_i) = \phi(q, \mathcal{N}_{c_i} \cup \mathcal{N}_{p_i}) \quad (3)$$

## 6 Experiments: Visual Paraphrasing Improves Image Captioning

The experimental configuration basically follows §4. We compare ASSOC<sub>para</sub>, the visual-paraphrase augmented approach, to the vanilla ASSOC approach. The image feature setting is the one with which the ASSOC approach performs best. Both approaches use the GIST+Tinyimage feature to prepare candidate captions, then use either the GIST or Tinyimage feature for reranking.

Table 8 shows that the ASSOC<sub>para</sub> approach significantly improves the vanilla ASSOC method under both automatic and human evaluation. As a reference, the first row shows the performance of the INSTANCE method (§4). The ASSOC method significantly improves over the INSTANCE method. On a similar vein, the ASSOC<sub>para</sub> method further improves over the ASSOC method, as automatic paraphrases provide a better visual neighborhood. This improvement is remarkable since the paraphrasing association is added automatically without any supervised training. This demonstrates the usefulness of the bipartite association structure of our corpus.

## 7 Image Captioning with Creativity

Naturally existing captions reflect everyday narratives, which in turn reflect figurative language use such as metaphor, simile, and personification. To gain better insights, one of the authors manually categorized a set of 1000 random captions. About 17%

are identified as figurative. Table 3 shows the distribution over different types of figurative captions.

**Creative Language Classifier:** Using the small set of labels described above, we train a simple binary classifier to identify captions with creative language.<sup>8</sup> Using this classifier, we can control the degree of literalness or creativity in generated captions. Based on 5-fold cross-validation, the classifier performs with 77% precision and 43% recall.

Importantly, a high-precision and low-recall classifier suffices our purpose. It is because in the context of creative captioning and creative paraphrasing presented below, we only need to detect *some* figurative captions, not *all*.

### 7.1 Creative Image Captioning

Given a query image  $q$ , we describe it with the most appropriate figurative caption. We propose the ASSOC<sub>creative</sub> approach that alters the ASSOC approach (§3) to return a *figurative* caption from the candidate pool, excluding *literal* captions.

### 7.2 Creative Visual Paraphrasing

Given a query image  $q$  and its *original* caption  $c$ , we rephrase  $c$  to a more creative and inspirational caption that still describes  $q$ . We use the PARA<sub>creative</sub> approach that changes our automatic visual paraphrasing algorithm (§5), by retrieving only figurative captions.

## 8 Experiments: Creative Image Captioning and Paraphrasing

### 8.1 Creative Captioning

We compare the ASSOC<sub>creative</sub> approach to the vanilla ASSOC approach. With the ASSOC approach, the top-rank caption is usually literal. Both approaches use the GIST+Tinyimage feature for preparing candidate captions, and the Tinyimage feature for reranking, which is the best setting for the ASSOC approach (§4).

Similarly to §4, we sample 200 test images from our corpus, and use AMT to compare two algorithms in terms of *visual relevance* and *creativity* separately. For creativity, we ask turkers to choose one

<sup>8</sup>We use a random forest classifier with features including words indicating reasoning (but, could, that), generality (never, always), caption length, abstract nouns (life, and hope), and whether the caption is a known idiom or quotation.

method	creativity	relevance
ASSOC	33%	41%
ASSOC <sub>creative</sub>	67%	59%

Table 9: Human evaluation for creative captioning: % of captions preferred by judges in pairwise comparisons

of the two captions that is more creative and inspirational than the other to describe each given test image. Results are shown in Table 9.

(1) *Creativity*. For 2/3 of the query images, captions produced by the ASSOC<sub>creative</sub> method are judged as more creative than those produced by the ASSOC method. This result indirectly validates that the figurativeness classifier has a reasonable precision to control the literalness of the system caption.

(2) *Visual relevance*. Interestingly, not only the captions from the ASSOC<sub>creative</sub> method are favored as creative, they are also judged as visually more relevant than those from the ASSOC method, despite that each figurative caption has lower neighborhood-based affinity than the literal counterpart. We conjecture that it is easier for human judges to be imaginative and draw visual relevance between the query image and figurative captions than the literal counterparts. This result also suggests that figurative language may be of practical use in image caption applications as a means to smooth the potentially brittle system output. Figure 5 shows example system output.

## 8.2 Creative Visual Paraphrasing

We test 200 images that are associated with *literal* captions as predicted by the figurativeness classifier. The PARA<sub>creative</sub> approach competes against two baselines: 1) the ORIGINAL captions, and 2) a text-only variant of the PARA<sub>caption</sub> approach sans visual processing: it randomly chooses a figurative caption that shares the largest number of physical-object nouns with the original caption, without looking at the query image. This is for evaluating the effect of visual context.

In addition to the evaluations as in §8.1, we also use a *multiple-choice* setting that allows a turker to choose zero to two captions that are visually relevant to the query image. See Table 10 for results, and Figure 5 for example outputs.

method	creativity	relevance	
		<i>single</i>	<i>multiple</i>
ORIGINAL	32%	80%	87%
PARA <sub>creative</sub>	68%	20%	60%
PARA <sub>caption</sub>	56%	47%	63%
PARA <sub>creative</sub>	44%	53%	74%

Table 10: Human eval for creative visual paraphrasing

**I. Comparing original captions with creative paraphrases (ORIGINAL vs. PARA<sub>creative</sub>):** The paraphrases are preferred over the original literal captions as more creative most of the time. As for the visual relevance, the original captions are favored over the paraphrases most of the time in the single-choice competition. However, when we use a multiple-choice setting, paraphrases has a reasonable relevance rate (60%), despite the simplicity of the algorithm. The fact that the original captions has a high relevance rate (87%) shows that in our corpus the captions have high visual relevance to their associated images most of the time.

**II. Creative paraphrasing with and without the visual context (PARA<sub>caption</sub> vs. PARA<sub>creative</sub>):** In terms of creativity, the PARA<sub>caption</sub> method is preferred over the PARA<sub>creative</sub> method. We conjecture that without conditioning on the visual content, PARA<sub>caption</sub> method tends to retrieve more unexpected captions that make turkers think they are more fun and creative. As for the visual relevance, by conditioning on the visual context given by query images, the PARA<sub>creative</sub> method significantly improves the visual relevance over the text-only counterpart, PARA<sub>caption</sub> method. This result highlights the pragmatic differences between visually-situated paraphrasing and text-based paraphrasing.

## 9 Related work

**Image-caption corpus:** Our work contributes to the line of research that makes use of internet web imagery and text (Ordonez et al., 2011; Berg et al., 2010) by detecting the visually relevant text (Dodge et al., 2012) and reducing the noise (Kuznetsova et al., 2013b; Kuznetsova et al., 2014). Compared to datasets with crowd-sourced captions (Hodosh et al., 2013; Lin et al., 2014), in which each image is annotated with several captions, our dataset presents several images for each caption, a subset of which also includes visually situated paraphrases. The as-







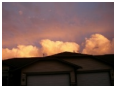







Creative Image Captioning		Creative Visual Paraphrasing	
< Good >	< Bad >	< Good >	< Bad >
 <ul style="list-style-type: none"> <li>-Hood under a full moon (*)</li> <li>-Mirror, mirror on the lake</li> </ul>	 <ul style="list-style-type: none"> <li>-Falling water(*)</li> <li>-Can you see the dogs</li> </ul>	 <ul style="list-style-type: none"> <li>-Bee on orange flowers(*)</li> <li>-When the flower looms, the bees come uninvited</li> </ul>	 <ul style="list-style-type: none"> <li>-long haired girl(*)</li> <li>-Diamonds are a girl's best friend</li> </ul>
 <ul style="list-style-type: none"> <li>-Sky on the way home(*)</li> <li>-Red sky at night, Shepherd's delight</li> </ul>	 <ul style="list-style-type: none"> <li>-City of lights (*)</li> <li>-Great balls of fire</li> </ul>	 <ul style="list-style-type: none"> <li>-Lights in cave(*)</li> <li>-There is a light that never goes out</li> </ul>	 <ul style="list-style-type: none"> <li>-Young roe deer(*)</li> <li>-The tree that looks like a deer</li> </ul>
 <ul style="list-style-type: none"> <li>-Sail on by (*)</li> <li>-Row, row, row your boat gently down the stream</li> </ul>	 <ul style="list-style-type: none"> <li>-Red Bean Pastries (*)</li> <li>-When life gives you lemons</li> </ul>	 <ul style="list-style-type: none"> <li>-Sky on the way home(*)</li> <li>-Go home, sky, you're drunk</li> </ul>	 <ul style="list-style-type: none"> <li>-The flight of the crane(*)</li> <li>-That's a crane</li> </ul>

Figure 5: Examples of creative captioning and creative visual paraphrasing. The left column shows good examples in blue, and the right column shows bad examples in red. The captions marked with \* are the original captions of the corresponding query images.

sociation structure of our dataset is analogous to that of ImageNet (Deng et al., 2009). Unlike ImageNet that is built for nouns (physical objects) listed under WordNet (Miller, 1995), our corpus is built for expressive phrases and full sentences and constructed without human curation. Our corpus has several unique properties to complement existing corpora. As explored in a very recent work of (Gong et al., 2014), we expect that it is possible to combine crowd-sourced and web-harvested datasets and achieve the best of both worlds.

**Image captioning:** Our work contributes to the increasing body of research on retrieval-based image captioning (Ordonez et al., 2011; Hodosh et al., 2013; Hodosh and Hockenmaier, 2013; Socher et al., 2014), by providing a new large-scale corpus with unique association structure between images and captions, by proposing an algorithm that exploits the structure, and by exploring two new dimensions: (i) visually situated paraphrasing (and its utility for retrieval-based image captioning), and (ii) creative image captioning.

**Paraphrasing:** Most previous studies in paraphrasing have focused exclusively on text, and the primary goal has been learning *semantic* equivalence of phrases that would be true out of context (e.g., (Barzilay and McKeown, 2001; Pang et al., 2003; Dolan et al., 2004; Ganitkevitch et al., 2013)), rather than targeting *situated* or *pragmatic* equivalence given a context. Emerging efforts began exploring paraphrases that are situated in video content (Chen and Dolan, 2011), news events (Zhang and Weld, 2013), and knowledge base (Berant and Liang, 2014). Our work is the first to introduce vi-

*sually situated paraphrasing* in which the task is to find paraphrases that are conditioned on both the input text as well as the visual context. (Chen and Dolan, 2011) collected situated paraphrases only through crowd sourcing, while we also explore automatic collection, and further test the quality of automatic paraphrases by using the learned paraphrases in an extrinsic evaluation setting.

**Figurative language:** There has been substantial work for detecting and interpreting figurative language (Shutova, 2010; Li et al., 2013; Kuznetsova et al., 2013a; Tsvetkov et al., 2014), while relatively less work on *generating* creative or figurative language (Veale, 2011; Ozbal and Strapparava, 2012). We probe data-driven approaches to creative language generation in the context of image captioning.

## 10 Conclusion

To conclude, we have provided insights into making a better use of multimodal web data in the wild, resulting in a large-scale corpus, *Deja Image-Captions*, with several unique properties to complement datasets with crowdsourced captions. To validate the usefulness of the corpus, we proposed new image captioning algorithms using the associative structure, which we extended to several related tasks ranging from visually situated paraphrasing to enhanced image captioning. In the process we have also explored several new tasks: visually situated paraphrasing, creative image captioning, and creative caption paraphrasing.

**Acknowledgement** The research is supported in part by NSF Award IIS 1447549 and IIS 1408287.

## References

- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics.
- Jonathan Berant and Percy Liang. 2014. Semantic Parsing via Paraphrasing. In *Association for Computational Linguistics (ACL)*.
- Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *ECCV 2010*, pages 663–676. Springer.
- Oren Boiman, Eli Shechtman, and Michal Irani. 2008. In defense of Nearest-Neighbor based image classification. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pages 1–8, June.
- David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.
- Thomas Deselaers and Vittorio Ferrari. 2011. Visual and semantic similarity in imagenet. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1777–1784. IEEE.
- Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and others. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772. Association for Computational Linguistics.
- Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.
- Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 452–457.
- Song Feng, Sujith Ravi, Ravi Kumar, Polina Kuznetsova, Wei Liu, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2015. Refer-to-as Relations as Semantic Knowledge. In *AAAI Conference on Artificial Intelligence*.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.
- Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In *ECCV 2014*, pages 529–545. Springer.
- David Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.
- Micah Hodosh and Julia Hockenmaier. 2013. Sentence-based image description with scalable, explicit models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 294–300.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47(1):853–899.
- Polina Kuznetsova, Jianfu Chen, and Yejin Choi. 2013a. Understanding and Quantifying Creativity in Lexical Composition. In *EMNLP*, pages 1246–1258.
- Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2013b. Generalizing Image Captions for Image-Text Parallel Corpus. In *ACL (2)*, pages 790–796.
- Polina Kuznetsova, Vicente Ordonez, Tamara Berg, and Yejin Choi. 2014. TreeTalk: Composition and Compression of Trees for Image Descriptions. *Transactions of the Association for Computational Linguistics*.
- Hongsong Li, Kenny Q. Zhu, and Haixun Wang. 2013. Data-Driven Metaphor Recognition and Explanation. *TACL*, 1:379–390.
- 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, Zürich*.
- Rebecca Mason and Eugene Charniak. 2014. Nonparametric Method for Data-driven Image Captioning. In *NAACL*.

- Mary Meeker. 2014. *Internet Trends 2014*.
- George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.
- Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, volume 1, page 4.
- Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. From large scale image categorization to entry-level categories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2768–2775. IEEE.
- Gozde Ozbal and Carlo Strapparava. 2012. A Computational Approach to the Automation of Creative Naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 703–711, Jeju Island, Korea, July. Association for Computational Linguistics.
- Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 102–109. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing 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.
- Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon’s Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, CSLDAMT ’10*, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Nikhil Rasiwasia, Pedro J. Moreno, and Nuno Vasconcelos. 2007. Bridging the gap: Query by semantic example. *Multimedia, IEEE Transactions on*, 9(5):923–938.
- Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A New Approach to Cross-modal Multimedia Retrieval. In *Proceedings of the International Conference on Multimedia, MM ’10*, pages 251–260, New York, NY, USA. ACM.
- Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL ’10*, pages 688–697, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.
- Antonio Torralba, Robert Fergus, and William T. Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958–1970.
- Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of ACL*.
- Tony Veale. 2011. Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 278–287, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.
- Congle Zhang and Daniel S Weld. 2013. Harvesting Parallel News Streams to Generate Paraphrases of Event Relations. In *EMNLP*, pages 1776–1786.