Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models

Bryan A. Plummer[†] Liwei Wang[†] Chris M. Cervantes[†] Juan C. Caicedo[*]
Julia Hockenmaier[†] Svetlana Lazebnik[†]
[†]Univ. of Illinois at Urbana-Champaign [*]Fundación Univ. Konrad Lorenz
[bplumme2,lwang97,ccervan2,juliahmr,slazebni]@illinois.edu
juanc.caicedor@konradlorenz.edu.co

Abstract

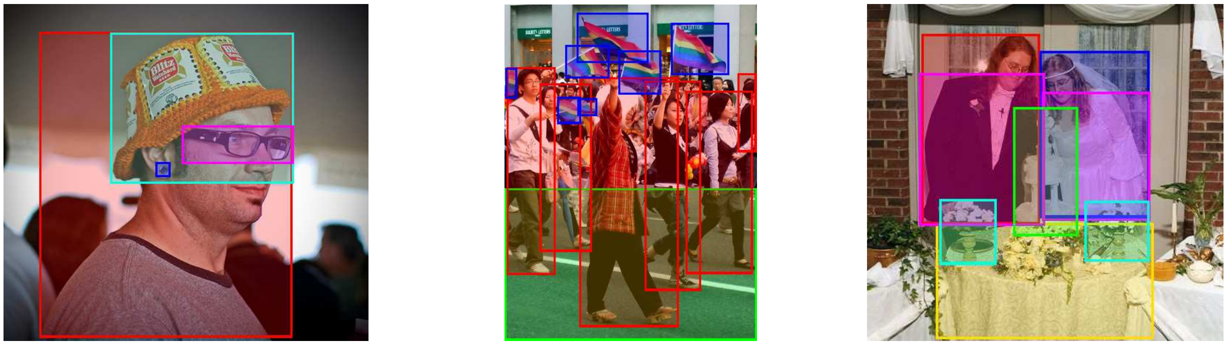
The Flickr30k dataset has become a standard benchmark for sentence-based image description. This paper presents Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains linking mentions of the same entities in images, as well as 276k manually annotated bounding boxes corresponding to each entity. Such annotation is essential for continued progress in automatic image description and grounded language understanding. We present experiments demonstrating the usefulness of our annotations for text-to-image reference resolution, or the task of localizing textual entity mentions in an image, and for bidirectional image-sentence retrieval. These experiments confirm that we can further improve the accuracy of state-of-the-art retrieval methods by training with explicit region-to-phrase correspondence, but at the same time, they show that accurately inferring this correspondence given an image and a caption remains really challenging.

1. Introduction

We are interested in understanding language grounded in visual content, and using language to better interpret images. The task of sentence-based image description, which combines both of these goals, has received a lot of attention in both computer vision and natural language processing communities [1, 4, 6, 7, 12, 16, 19, 20, 22, 23, 25, 28, 37, 39]. Existing datasets for this task pair each image with one or more captions [11, 12, 24, 31, 40]. Unfortunately, none of these datasets provide an explicit grounding of where the entities mentioned in the captions appear in the image. As a consequence, most approaches to automatic image description either learn global associations between images and sentences without any explicit attempt to detect or localize the mentioned entities [1, 4, 12, 19, 20, 23, 25, 37], or rely on detectors that were trained for different purposes [7, 22, 28, 39]. A number of recent works have taken a more conceptually satisfying approach of inducing mappings of image regions to words or phrases in the captions [6, 17, 16, 38], but have had to treat these mappings as latent. It is reasonable to believe the accuracy of the latter methods would be enhanced by having explicit supervision at training time. At test time, ground-truth region-to-text correspondence could help evaluate how accurately methods associate phrases with specific image locations. Indeed, there is evidence that state-of-the-art caption generation methods tend to reproduce generic captions from the training data and do not perform well on compositionally novel images [2]. To overcome such weaknesses and develop richer compositional image-sentence models, we need large-scale supervision and new benchmarks.

The main contribution of this paper is providing the first large-scale comprehensive dataset of region-to-phrase correspondences for image description. We build on the Flickr30k dataset [40], a popular benchmark for caption generation and retrieval [1, 4, 6, 10, 17, 16, 19, 20, 23, 25, 37, 38]. Flickr30k contains 31,783 images focusing mainly on people and animals, and 158,915 English captions (five per image). In this work, we introduce Flickr30k Entities, which augments the original dataset by identifying which mentions among the captions of the same image refer to the same set of entities, resulting in 244,035 *coreference chains*, and which image regions depict the mentioned entities, resulting in 275,775 bounding boxes. Figure 1 illustrates the structure of our annotations on three sample images.

The largest existing dataset with both captions and region-level annotations is Microsoft Common Objects in Context (MSCOCO) [24], containing over 300k images with over 2.5m labeled object instances from 80 pre-defined categories, and five captions per image. However, the MSCOCO region-level annotations were produced independently from its captions, and phrases in the captions are not linked to these regions or to each other in any way. Thus, while MSCOCO is an order of magnitude larger, unlike our



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .

A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Figure 1: Example annotations from our dataset. In each group of captions describing the same image, coreferent mentions (*coreference chains*) and their corresponding bounding boxes are marked with the same color. On the left, each chain points to a single entity (bounding box). Scenes and events like "outside" or "parade" have no box. In the middle example, the people (red) and flags (blue) chains point to multiple boxes each. On the right, blue phrases refer to the bride, and red phrases refer to the groom. The dark purple phrases ("a couple") refer to both of these entities, and their corresponding bounding boxes are identical to the red and blue ones.

dataset, it has neither cross-caption coreference information nor explicit groundings of textual mentions to regions.

Our dataset is also related to ReferIt [18], which augments the IAPR-TC dataset [11] of 20k photographs with 130k isolated entity descriptions for 97k objects (image regions), only 10k of which have more than a single description. While some of the ReferIt descriptions contain spatial relations to other objects, they typically do so only if that is necessary to uniquely identify the object in the image. The average number of identified objects per image in our dataset (8.9 bounding boxes) is significantly higher than in the ReferIt dataset (4.8), and is on par with MSCOCO (7.7).

Johnson et al. [15] is another notable work concerned with grounding of semantic scene descriptions to image regions. However, instead of natural language, it proposes a formal *scene graph* representation that encapsulates all entities, attributes and relations in an image, together with a dataset of scene graphs and their groundings for 5k images. While these graphs are more dense and detailed than our annotations (on average, each image has 13.8 objects, 18.9 attributes and 21.9 relations), such an exhaustive annotation scheme makes it difficult to identify salient content for natural language communication, and it is unclear how it can scale to larger numbers of images.

Section 2 describes our crowdsourcing protocol, which consists of two major stages, coreference resolution and bounding box drawing. Each stage is split up into smaller tasks to ensure both efficiency and quality. Our annotations also enable us to conduct quantitative evaluation for new benchmark tasks, such as text-to-image reference resolution. Section 3 presents baseline results for this task, demonstrating that even with state-of-the-art text-to-image embeddings, accurately localizing specific entities in an image remains challenging. Finally, in Section 4, we show how our region-phrase annotations can help to improve performance on the established task of bidirectional (image-to-sentence and sentence-to-image) retrieval.

2. Annotation Process

As can be seen from Figure 1, our desired annotations are highly structured. They also vary in complexity from image to image, since images vary in the numbers of clearly distinguishable entities they contain, and sentences vary in the extent of their detail. Further, there are ambiguities involved in identifying whether two mentions refer to the same entity or set of entities, how many boxes (if any) these entities require, and whether these boxes are of sufficiently high quality. Due to this intrinsic subtlety of our task, compounded by the unreliability of crowdsourced judgments, we developed a multi-stage pipeline of simpler atomic tasks to ensure high-quality final annotations.

Our pipeline consists of two stages: **coreference resolution**, or forming coreference chains that refer to the same entities (Section 2.1), and **bounding box annotation** for the resulting chains (Section 2.2). This workflow provides two advantages: first, identifying coreferent mentions helps reduce redundancy and save box-drawing effort; and second, coreference annotation is intrinsically valuable, e.g., for training cross-caption coreference models [13]. More details on the interfaces used to collect annotations on Amazon's Mechanical Turk (AMT) are provided in the supplementary material.

2.1. Coreference Resolution

We rely on the chunking information given in the Flickr30k captions [40] to identify potential entity mentions. With the exception of personal pronouns (*he, she, they*) and a small list of frequent non-visual terms (*background*, *air*), we assume that any noun-phrase (NP) chunk is a potential entity mention. NP chunks are short (avg. 2.35 words), non-recursive phrases (e.g., the complex NP *[[a man] in [an orange hat]]* is split into two chunks). Mentions may refer to single entities (*a dog*); regions of "stuff" (*grass*); multiple distinct entities (*two men*, *flags*, *football players*); groups of entities that may not be easily identified as individuals (*a crowd*, *a pile of oranges*); or even the entire scene (*the park*). Finally, some NP chunks may not refer to any physical entities (*wedding reception*, *a trick*, *fun*).

Once we have our candidate mentions from the sentences corresponding to the same image, we need to identify which ones refer to the same set of entities. Since each caption is a single, relatively short sentence, pronouns (*he*, *she*, *they*) are relatively rare in this dataset. Therefore, unlike in standard coreference resolution in running text [34], which can be beneficial for identifying all mentions of people in movie scripts [30], we ignore anaphoric references between pronouns and their antecedents and focus on cross-caption coreference resolution [13]. Like standard coreference resolution, our task partitions the set of mentions M in a document (here, the five captions of one image), into subsets of equivalent mentions such that all mentions in the same subset $c \in C$ refer to the same set of entities. In keeping with standard terminology, we refer to each such set or cluster of mentions $c \subset M$ as a coreference chain.

Binary Coreference Link Annotation. Since the task of constructing an entire coreference chain from scratch is cognitively complex and error-prone, we broke it down into a simpler task to collect binary coreference links between pairs of mentions. A coreference link between mentions m and m' indicates that m and m' refer to the same set of entities. In the manual annotation process, workers are shown an image and the two captions from which m and m' originate. The workers are asked whether these mentions refer to the same entity. If a worker indicates that the mentions are coreferent, we add a link between m and m'. Given a set of mentions M for an images, manual annotation of all $O(|M|^2)$ pairwise links is prohibitively costly. But since M typically contains multiple mentions that refer to the same set of entities, the number of coreference chains is bounded by, and typically much smaller than, $|M|$. This allows us to reduce the number of links that need to be annotated to $O(|M||C|)$ by leveraging the transitivity of the coreference relation [26]. Given a set of identified coreference chains C and a new mention m that has not been annotated for coreference yet, we only have to ask for links between m and one mention from each element of C. If m is not coreferent with any of these mentions, it refers to a new entity whose coreference chain is initialized and added to C.

In the worst case, each entity has only one mention requiring annotation of all $|M|^2$ possible links. But in practice, most images have more mentions than coreference chains (in our final dataset, each image has an average of 16.6 mentions and 7.8 coreference chains). We further reduce the number of required annotations with two simplifying assumptions. First, we assume that mentions from the same captions cannot be coreferent, as it would be unlikely for a caption to contain two non-pronominal mentions to the same set of entities. Second, we categorize each mention into eight coarse-grained types using manually constructed dictionaries (people, body parts, animals, clothing/color[1], instruments, vehicles, scene, and other), and assume mentions belonging to different categories cannot be coreferent.

Coreference Chain Verification. To handle errors introduced by the coreference link annotation, we verify the accuracy of all chains that contain more than a single mention. Although false negatives (missing coreference links) lead to an oversegmentation of entities that increases the time required to draw boxes for each set of entities, we can identify this redundancy post-hoc since the associated boxes are highly likely to have significant overlap (see Section 2.3 for details on box merging). False positives (spurious coreference links) are more harmful for our purposes, since they may result in mentions being associated with incorrect entities or image regions. We use a Coreference Chain Verification task to detect these false positive coreference links. Here, workers are shown the mentions that belong to the same coreference chain and asked whether all the mentions refer to the same set of entities. If the worker answers True, the chain is kept as-is. If a worker answers False, that chain is broken into subsets of mentions that share the same head noun (the last word in a chunk).

2.2. Bounding Box Annotations

The workflow to collect bounding box annotations is broken down similarly to Su *et al.* [36], and consists of four separate AMT tasks, discussed below: (1) Box Requirement, (2) Box Drawing, (3) Box Quality, and (4) Box Coverage. In each task, workers are shown an image and a caption in which a representative mention for one coreference chain is highlighted. We use the longest mention in each chain, since we assume that it is the most specific.

[1] In Flickr30k, NP chunks that only consist of a color term are often used to refer to clothing, e.g. *man in blue*.

Box Requirement. First, we determine if the entities a representative mention refers to require boxes to be drawn. A mention does not require boxes if it refers to the entire scene (*in [the park]*), to physical entities that are not in the image (*pose for [the camera]*), or to an action or abstract entity (*perform [a trick]*). Given an image and a caption with a highlighted mention, we ask workers if (1) at least one box can be drawn (2) the mention refers to a scene or place or (3) no box can be drawn.

If the worker determines that at least one box can be drawn, the coreference chain proceeds to the Box Drawing task (below). Otherwise, we ask for a second and sometimes a third Box Requirement judgment to obtain agreement between two workers. If the majority agrees that no box needs to be drawn, the coreference chain is marked as "non-visual" and leaves the bounding box annotation workflow. After preliminary analysis, we determined that coreference chains with mentions from the people, clothing, and body parts categories so frequently required boxes that they immediately proceeded to the Box Drawing task, skipping the Box Requirement task altogether.

Box Drawing. In this task, we collect bounding boxes for a mention. The key source of difficulty here is due to mentions that refer to multiple entities. Our annotation instructions specify that we expect individual boxes around each entity if these can be clearly identified (e.g., *two people* would require two boxes). But if individual elements of a group cannot be distinguished (*a crowd of people*), a single box may be drawn around the group. We show workers all previously drawn boxes for the representative mention (if they exist), and ask them to draw one new box around one entity referred to by the mention, or to indicate that no further boxes are required.

If the worker adds a box, the mention-box pair proceeds to the Box Quality task. If the worker indicates that no boxes are required, the mention accrues a "no box needed" judgment. The mention is then returned to Box Requirement if it has no boxes associated with it. Otherwise, the mention is sent to Box Coverage.

Box Quality. For each newly drawn we ask a worker whether the box is good. Since we want to avoid redundant boxes, we also show all previously drawn boxes for the same mention. Good boxes are tightly drawn around the entire entity a mention refers to which no other box already covers. When mentions refer to multiple entities that can be clearly distinguished, these must be associated with individual boxes. If the worker marks the box as 'bad', it is discarded and the mention is returned to the Box Drawing task. If the worker marks the box as 'good', the mention proceeds to the Box Coverage task to determine whether additional boxes are necessary.

Box Coverage. In this step, workers are shown the boxes that have been drawn for a mention, and asked if all required boxes are present for that mention If the initial judgment says that more boxes are needed, the mention is immediately sent back to Box Drawing. Otherwise, we require a second worker to verify the decision that all boxes have been drawn. If the second worker disagrees, we collect a third judgment to break the tie, and either send the mention back to Box Drawing, or assume all boxes have been drawn.

2.3. Quality Control

Identifying Trusted Workers. Since annotation quality on AMT is highly variable [35, 31], we only allow workers who have completed at least 500 previous HITs with 95% accuracy, have successfully completed a corresponding qualification test for each of our six tasks, and have sufficient accuracy on their first 30 items. For Box Drawing, Boxes have to pass Box Quality. For the binary tasks, we use verification questions for which we know the answer. For the binary tasks, they then have to maintain high accuracy on the 2% of items that are also verification questions.

Additional Review. At the end of the crowdsourcing process, we identified roughly 4k entities that required additional review. This included some chunking errors that came to our attention (e.g., through worker comments), as well as chains that cycled repeatedly through the Box Requirement or Box Coverage task, indicating disagreement among the workers. Images with the most serious errors were manually reviewed by the authors.

Box and Coreference Chain Merging. As discussed in Section 2.1, coreference chains may be fragmented due to missed links (false negative judgments). Additionally, if an image contains more than one entity of the same type, its coreference chains may overlap or intersect (e.g., *a bride* and *a couple* from Figure 1). Since Box Drawing operates over coreference chains, it results in redundant boxes for such cases. We remove this redundancy by merging boxes with IOU scores of at least 0.8 (or 0.9 for "other"). This process has some restrictions (e.g. clothing and people boxes cannot be merged). Afterwards, we merge any coreference chains that point to the exact same set of boxes.

Error Analysis. Errors present in our dataset mostly fall under two categories: chunking and coreference errors. Chunking errors occur when the automated tools made a mistake when identifying mentions in caption text. Coreference errors occur when AMT workers made a bad judgment when building coreference chains. An analysis using a combination of automated tools and manual methods identified chunking errors in less than 1% of the dataset's mentions and coreference errors in less than 1% of the datasets chains. Since, on average, there are over 16 mentions and 7 chains per image, there is an error of some kind in around 8% of our images.

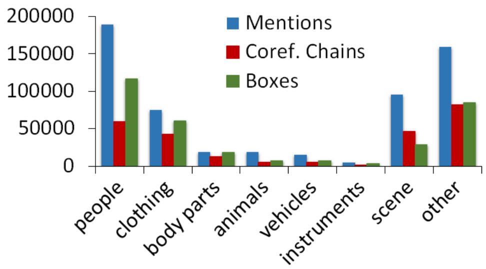


Figure 2: The total number of coreference chains, mentions, and bounding boxes per type.

Type	#Chains	Mentions/Chain	Boxes/Chain
people	59766	3.17	1.95
clothing	42380	1.76	1.44
body parts	12809	1.50	1.42
animals	5086	3.63	1.44
vehicles	5561	2.77	1.21
instruments	1827	2.85	1.61
scene	46919	2.03	0.62
other	82098	1.94	1.04
total	244035	2.10	1.13

Table 1: Coreference chain statistics. The number of mentions per chain indicates how salient an entity is. The number of boxes per chain indicates how many distinct entities it refers to.

2.4. Dataset Statistics

Our annotation process has identified 513,644 entity or scene mentions in the 158,915 Flickr30k captions (3.2 per caption), and these have been linked into 244,035 coreference chains (7.7 per image). The box drawing process has yielded 275,775 bounding boxes in the 31,783 images (8.7 per image). Figure 2 shows the distribution of coreference chains, mentions, and bounding boxes across types, and Table 1 shows additional coreference chain statistics. 48.6% of the chains contain more than a single mention. The number of mentions per chain varies significantly across entity types, with salient entities such as people or animals being mentioned more frequently than e.g clothing or body parts. Aggregating across all five captions, people are mentioned in 94.2% of the images, animals in 12.0%, clothing and body parts in 69.9% and 28.0%, vehicles and instruments in 13.8% and 4.3%, while other objects are mentioned in 91.8% of the images. The scene is mentioned in 79.7% of images. 59.1% of the coreference chains are associated with a single bounding box, 20.0% with multiple bounding boxes, and 20.9% with no bounding box, but there is again some wide variety across entity types. The people category has significantly more boxes than chains (116k boxes for 60k chains) suggesting that many of these chains describe multiple individuals (*a family*, *a group of people*, etc.).

Training Set	Size
All NP chunks	423,134
Resampled, $N = 1$	70,759
Resampled, $N = 10$	137,133

Table 4: Training set sizes for our experiments.

3. Text-to-Image Reference Resolution

As a key application for our annotations, we consider the task of *text-to-image reference resolution*, i.e., grounding or localizing textual mentions of entities in an image. To our knowledge, Kong *et al.* [21] is the only work that deals directly with this task, but it is focused on using sentences to help with 3D parsing of RGB-D images. Up to now, in the absence of the kinds of annotations provided by Flickr30k Entities, it has not been possible to use general text-to-image reference resolution as a benchmark task for image description.

Given an image and a sentence that describes it, we want to predict a bounding box for each entity mention from that sentence. This task is akin to object detection and can be evaluated in a similar way. However, training a separate detector for each unique noun phrase is not a promising solution since a large number of phrases are very rare, and different phrases may refer to similar entities (e.g. *infant* and *baby*). Instead, as a baseline approach, we learn an embedding of region and phrase features to a shared latent space and use distance in that space to score image regions.

3.1. Region-Phrase CCA Model

We want to learn a shared semantic space which would allow us to associate phrases in our sentences to image regions. This can be done in various ways, from recurrent neural networks [16, 19, 25] to Canonical Correlation Analysis (CCA) [10, 20]. Even though CCA is a classic technique [14], Klein *et al.* [20] have recently used it to achieve remarkable results for image-sentence retrieval. Key to their performance are state-of-the-art features for images (deep activation features from VGG net [32]) and text (Fisher vector pooling [29] on word2vec vectors [27] of individual words). Due to the simplicity, high accuracy, and speed of this model (on Flick30K, training of CCA only takes a couple of minutes, while recurrent neural networks may need tens of hours), we adopt it as our baseline.

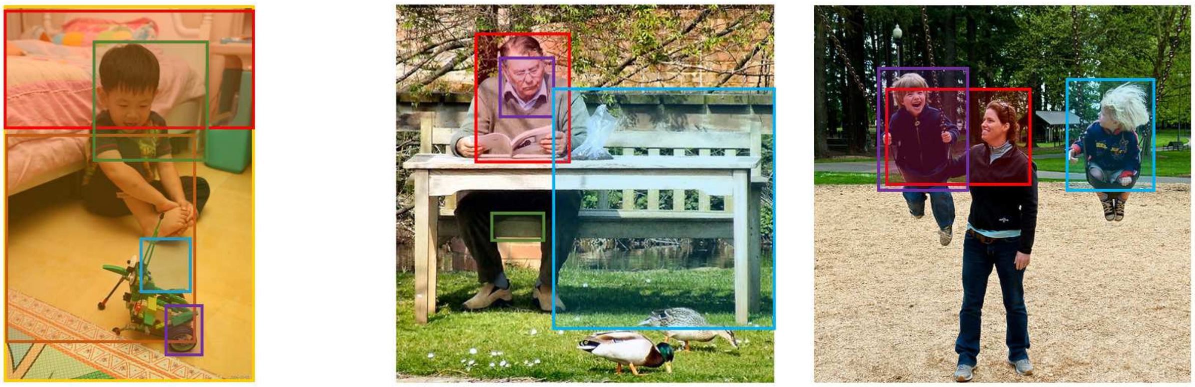
We generally follow the implementation details in [20]. Given an image region, we represent it using 4096-dimensional activations from the 19-layer VGG model (unlike whole-image features of [20], which are averaged over ten different crops, our region-level features are computed from a single crop, which we have found to give better results). Given a phrase, we represent each word with a 300-D word2vec feature. Then we construct a Fisher Vec-

	people	clothing	bodyparts	animals	vehicles	instruments	scene	other	mAP/overall
#Instances	5656	2306	523	518	400	162	1619	3374	14558
AP-NMS	13.16	11.48	4.85	13.84	13.67	11.42	10.86	10.50	11.22
R@1	29.58	24.20	10.52	33.40	34.75	35.80	20.20	20.75	25.30
R@10	71.25	52.99	29.83	66.99	76.75	61.11	57.44	47.24	59.66
R@100	89.36	66.48	39.39	84.56	91.00	69.75	75.05	67.40	76.91

Table 2: Localization performance using our resampled ($N = 10$) CCA model to rank 100 object proposals per image. We report average precision after nonmaximum suppression (AP-NMS) for different training data and Recall@K (see Section 3 for details).

	man	woman	boy	girl	person	people	dog	two men	street	young boy	little girl	two people
#Instances	1100	599	140	135	132	119	116	112	101	97	84	84
AP-NMS	9.52	15.89	13.71	6.76	7.02	10.53	19.00	28.96	3.05	10.16	11.88	7.68
R@1	32.30	35.18	33.09	32.59	30.30	11.21	33.04	29.52	23.76	36.17	39.76	28.05
R@100	86.91	90.82	92.86	80.74	81.06	83.19	95.69	78.57	66.34	92.78	84.52	86.90
	water	child	little boy	two women	table	ball	hat	white shirt	young girl	young man	crowd	black shirt
#Instances	83	82	69	65	63	62	61	57	57	54	53	50
AP-NMS	14.97	4.37	5.82	11.33	14.54	9.09	16.36	11.13	20.98	11.48	13.45	14.10
R@1	35.37	9.76	48.53	30.16	24.19	40.98	38.98	26.32	43.64	48.15	24.53	8.00
R@100	65.06	91.46	88.41	93.85	74.60	43.55	62.30	68.42	92.98	87.04	88.68	68.00

Table 3: Localization performance for the 24 most common phrases in the test set.



A small Asian boy [0.45] is sitting on the floor [0.82] of a bedroom [0.87] being entertained and smiling at a lego toy [0.77] that looks like a bug [0.87] on wheels [0.81] .

A man [0.77] with parted hair [0.91] and wearing glasses [0.89] is seated outdoors on a bench [0.74] where he is reading .

A woman [0.81] pushes a child [0.67] on a swing [0.44] while another swinging child [0.67] looks on .

Figure 3: Example localization results. For each image and reference sentence, phrases and top matching regions are shown in the same color. The CCA matching score is given in brackets after each phrase (low scores are better). See Section 3.2 for discussion.

tor codebook with 30 centers using both first and second order information, resulting in phrase features of dimensionality $300 \times 30 \times 2 = 18000$. Due to memory constraints, we only use the Hybrid Gaussian-Laplacian Mixture Model (HGLMM) for our experiments rather than the combined HGLMM+GMM which reported the best performance in [20]. We use the normalized CCA formulation of [9], where we scale the columns of the CCA projection matrices by the eigenvalues and normalize feature vectors projected by these matrices to unit length. In the resulting space, we use cosine distance to rank image regions given a phrase.

3.2. Phrase Localization Evaluation

As explained above, using our CCA model we can compute a score for each entity mention from a given sentence to each image region. To obtain candidate image regions, we take the top 100 proposals returned by the EdgeBox method [41]. Following [10, 16, 20, 25], we use 29,783 images for training, 1,000 for validation, and 1,000 for testing. Our split is the same as in [10]. At test time, we use the Stanford parser [33] to identify NP chunks in the test captions, and attempt to localize phrases that exactly match a ground truth phrase. If no exact match is present, no prediction is made, which results in a penalty during evaluation.

We evaluate localization performance using recall and average precision. First, we treat phrase-region reference resolution as a retrieval problem with the phrase as the query and the proposals from the input image as the database to search over. For this setup, we report Recall@K ($K = 1, 10$), or the percentage of queries for which a correct match has rank of at most K (we deem a region to be a correct match if it has IOU ≥ 0.5 with the ground truth bounding box for that phrase). Since we use 100 proposals per image, reporting Recall@100 gives an upper bound on localization performance given our region proposals. Further, since we get a ranking of regions from test set images for each unique phrase from the test sentences, we can evaluate our output using standard object detection methodology [5]. To this end, we report average precision following non-maximum suppression of predicted boxes (AP-NMS).

As a first attempt, we train a CCA embedding on the set of all ground-truth phrase-region pairs from the dataset (for phrases associated with multiple regions, we merge them into a single bounding box). The resulting model gets 8.69 mean average precision (mAP) over our phrase types. Part of the reason for the low performance is that the distribution of region counts for different NP chunks in this training set is very unbalanced: a few NP chunks, like *a man*, are extremely common, while others, like *tattooed, shirtless young man*, occur quite rarely. We found we can alleviate this problem by keeping at most N randomly selected exemplars for each phrase. We get our best mAP of 11.22 by resampling the dataset with $N = 10$ (for $N = 1$, the mAP is only 7.70). The sizes of the corresponding training sets are listed in the first three lines of Table 4. A breakdown of the performance over phrase types using both evaluation metrics with our best model is found in Table 2. Table 3 shows performance for the 24 most frequent phrases.

While our average precision is poor, this can be partially explained by the fact that the CCA model is better suited for retrieval than for localization. For many phrase types, it is possible to get a good CCA score for a region in which the corresponding object is poorly localized. For example, to get a good score for "person," it is not necessary to have a tight bounding box that encloses the entire visible extent of the person in the picture (qualitative results in Figure 3 are indicative). We also obtain low recall on smaller entity types (*hat*, *ball*) due to a lack of good object proposals for them. On the other hand, not shown in Table 3 are 436 phrases that have AP of 100 (e.g., *cow*, *goalie*, *rugs*), but occur seven or fewer times in our test set.

Figure 3 shows example results for three image-sentence pairs, and helps to illustrate the challenges inherent in text-to-image reference. In the left example, we find the boy, the toy's wheels, and associate a large region containing the bed with "bedroom." In the center example we find a plausible box for the man, but the box for *glasses* lands on the bench beneath him. To deal with such errors, incorporating spatial constraints into our localization model will be crucial. The right example is especially tricky. Here, our model finds the same box for each child, while finding the swing for the correct child. In order to properly interpret the sentence cues, we need to determine from image evidence which child is being pushed by the woman and which one looks on, and this is likely to require sophisticated interaction models.

4. Image-Sentence Retrieval

Now that we have shown how our annotations can be used to train models and establish baselines for a new task, we would also like to demonstrate their usefulness for the standard tasks of sentence-to-image and image-to-sentence retrieval.

We use the standard protocol for evaluation: given the 1,000 images and 5,000 corresponding sentences in the test set, we use the images to retrieve the sentences and vice versa, and report performance as Recall@K, or the percentage of queries for which at least one correct ground truth match was ranked among the top K matches.

4.1. Region-Phrase Correspondences for Training

The results in Section 3 came from a CCA model trained on regions and phrases. Now we train a CCA model on whole-image and whole-sentence features using the implementation details in Section 3.1 (following [20], we average the whole-image features over ten crops for best performance). Table 5(b) shows the performance of this model on our test set. As can be seen from Table 5(a), this simple baseline outperforms more complex RNN-based models of [16, 19, 25]. Next, we want to leverage our region-phrase annotations to learn a better embedding for image-sentence retrieval.

An obvious way to train a model combining image-sentence correspondences with region-phrase correspondences would be to simply take the union of the image-sentence and region-phrase training sets. However, our image-level and region-level features are computed in different ways and have different statistics. In particular, because the image-level features are averaged over multiple crops while the region-level features are not, the latter are much sparser. The features for the sentences are also unlike those of phrases due to the differences in the number and kinds of words present in the two types of data. Thus, it is inappropriate to combine the image-sentence and region-phrase correspondences into a single training set. Instead, we adopt the *Stacked Auxiliary Embedding* (SAE) method of Gong *et al.* [10] where embedded features learned from auxiliary sources are concatenated with the original features to form a stacked representation. In the case of Gong *et al.* [10], the auxiliary data came from Flickr images and

	Methods on Flickr30K	Image Annotation			Image Search		
		R@1	R@5	R@10	R@1	R@5	R@10
(a) State of the art	BRNN [16]	22.2%	48.2%	61.4%	15.2%	37.7%	50.5%
	MNLM [19]	23.0%	50.7%	62.9%	16.8%	42.0%	56.5%
	m-RNN [25]	35.4%	63.8%	73.7%	22.8%	50.7%	63.1%
	GMM+HGLMM FV [20]	33.3%	62.0%	74.7%	25.6%	53.2%	66.8%
(b) Whole image-sentence CCA	HGLMM FV	36.5%	62.2%	73.3%	24.7%	53.4%	66.8%
(c) Combined image-sentence and region-phrase, $N = 10$ CCA	SAE	**38.3%**	**63.7%**	**75.5%**	**25.4%**	**54.2%**	**67.4%**
	Weighted Distance	**39.1%**	**64.8%**	**76.4%**	**26.9%**	**56.2%**	**69.7%**

Table 5: Bidirectional retrieval results. Image Annotation refers to using images to retrieve sentences, and Image Search refers to using sentences to retrieve images. The numbers in (a) come from published papers, and the numbers in (b) are from our own reproduction of the results of [20] using their code. See Section 4 for additional details.

unstructured tags, while we use our region-phrase pairs instead. As shown in the first line Table 5(c), the resulting model gives us an improvement of just over 1% over the image-sentence model. This demonstrates in principle that adding region- and phrase-level correspondences can help to train better bidirectional retrieval models. The demonstration is not fully satisfactory, since SAE does not care about the nature of the auxiliary data and does not take into account the actual structure and relationships between whole images and sentences and their parts, but it helps to establish a strong baseline and indicates that further research is likely to be fruitful.

4.2. Region-Phrase Correspondences for Retrieval

The CCA models of Section 4.1 are trained with the help of region-phrase correspondence. However, at test time, they are still used for global image-sentence retrieval, without attempting to match regions in a query image and phrases in a candidate matching sentence. To address this limitation, we can additionally draw on the region-phrase models from Section 3. Given a test sentence that we want to rank with respect to an image, for each phrase feature p_i from that sentence, we obtain the top-ranked candidate region $r(p_i)$. Then we compute the following distance between the sentence defined as a set of phrases and the image defined as the set of regions:

$$D_{RP} = \frac{1}{L^{\gamma}} \sum_{i=1}^{L} ||p_i - r(p_i)||_2^2 \,, \qquad (1)$$

where the exponent $\gamma \geq 1$ is meant to lessen the penalty associated with matching images to sentences with a larger number of phrases. Such sentences tend to mention more details that are harder to localize and therefore receive larger phrase-to-region distance scores than matches from shorter, more generic sentences. Experimentally, we have found $\gamma = 1.5$ to produce the best results. Finally, we define a new image-sentence distance as

$$\hat{D}_{IS} = \alpha D_{IS} + (1 - \alpha) D_{RP} \,, \qquad (2)$$

where D_{IS} is the squared Euclidean distance between CCA-projected global image and sentence features.

The second line of Table 5(d) shows results of this weighted distance with $\alpha = 0.7$ (by itself, the performance of eq. (1) is very poor). Compared to just using D_{IS} for retrieval, we get a consistent improvement of around 2% for image search and a smaller gain for image annotation. Once again, this demonstrates in principle that retrieval can be improved by attempting to infer the correspondence between regions and phrases at test time, and more research is needed to fully realize the potential of this idea.

5. Conclusion

This paper has presented Flickr30k Entities, the first image description dataset that provides comprehensive ground-truth correspondence between regions in images and phrases in captions. Our annotations can be used to benchmark tasks like text-to-image reference resolution, for which up to now large-scale ground-truth information has been lacking. While methods for global image description have been improving rapidly, our experiments suggest that current models are still quite weak at grounding specific textual mentions in local image regions, and datasets like ours are needed to continue to make progress on the problem.

Because our dataset is densely annotated with multiple boxes per image linked to their textual mentions in a larger sentence context, it will also be a rich resource for learning models of multi-object spatial layout [7, 8, 22]. Other potential applications include training models for automatic cross-caption coreference [13] and distinguishing visual vs. non-visual text [3].

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grants No. 1053856, 1205627, 1405883, IIS-1228082, and CIF-1302438 as well as support from Xerox UAC and the Sloan Foundation. We thank the NVIDIA Corporation for the generous donation of the GPUs used for our experiments.

References

[1] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv:1411.5654*, 2014. 1

[2] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. In *arxiv.org:1505.01809*, 2015. 1

[3] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. D. III, A. C. Berg, and T. L. Berg. Detecting visual text. In *NAACL*, 2012. 8

[4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv:1411.4389*, 2014. 1

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 7

[6] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. *arXiv:1411.4952*, 2014. 1

[7] A. Farhadi, S. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*. 2010. 1, 8

[8] S. Fidler, A. Sharma, and R. Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013. 8

[9] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014. 6

[10] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014. 1, 5, 6, 7

[11] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006. 1, 2

[12] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 2013. 1

[13] M. Hodosh, P. Young, C. Rashtchian, and J. Hockenmaier. Cross-caption coreference resolution for automatic image understanding. In *CoNLL*, pages 162–171. ACL, 2010. 3, 8

[14] H. Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936. 5

[15] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv:1412.2306*, 2014. 1, 5, 6, 7, 8

[17] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 1

[18] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2

[19] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*, 2014. 1, 5, 7, 8

[20] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *CVPR*, 2015. 1, 5, 6, 7, 8

[21] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 5

[22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011. 1, 8

[23] R. Lebret, P. O. Pinheiro, and R. Collobert. Phrase-based image captioning. *arXiv:1502.03671*, 2015. 1

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1

[25] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:1412.6632*, 2014. 1, 5, 6, 7, 8

[26] J. F. McCarthy and W. G. Lehnert. Using decision trees for coreference resolution. *arXiv cmp-lg/9505043*, 1995. 3

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 5

[28] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2Text: Describing images using 1 million captioned photographs. *NIPS*, 2011. 1

[29] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 5

[30] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with ”their” names using coreference resolution. In *ECCV*, 2014. 3

[31] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's mechanical turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. ACL, 2010. 1, 4

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 5

[33] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing With Compositional Vector Grammars. In *ACL*, 2013. 6

[34] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001. 3

[35] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. *Internet Vision Workshop*, 2008. 4

[36] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *AAAI Technical Report, 4th Human Computation Workshop*, 2012. 3

[37] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv:1411.4555*, 2014. 1

[38] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*, 2015. 1

[39] B. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2T: Image parsing to text description. *Proc. IEEE*, 98(8):1485 – 1508, 2010. 1

[40] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 1, 3

[41] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 6