Automatic Concept Discovery from Parallel Text and Visual Corpora

Chen Sun
Univ. of Southern California
chensun@usc.edu

Chuang Gan*
Tsinghua University
ganchuang1990@gmail.com

Ram Nevatia
Univ. of Southern California
nevatia@usc.edu

Abstract

Humans connect language and vision to perceive the world. How to build a similar connection for computers? One possible way is via visual concepts, which are text terms that relate to visually discriminative entities. We propose an automatic visual concept discovery algorithm using parallel text and visual corpora; it filters text terms based on the visual discriminative power of the associated images, and groups them into concepts using visual and semantic similarities. We illustrate the applications of the discovered concepts using bidirectional image and sentence retrieval task and image tagging task, and show that the discovered concepts not only outperform several large sets of manually selected concepts significantly, but also achieves the state-of-the-art performance in the retrieval task.

1. Introduction

Language and vision are both important for us to understand the world. Humans are good at connecting the two modalities. Consider the sentence “A fluffy dog leaps to catch a ball”: we can all relate fluffy dog, dog leap and catch ball to the visual world and describe them in our own words easily. However, to enable a computer to do something similar, we need to first understand what to learn from the visual world, and how to relate them to the text world.

Visual concepts are a natural choice to serve as the basic unit to connect language and vision. A visual concept is a subset of human vocabulary which specifies a group of visual entities (e.g., fluffy dog, curly dog). We name the collection of visual concepts as a visual vocabulary. Computer vision researchers have long collected image examples of manually selected visual concepts, and used them to train concept detectors. For example, ImageNet [6] selects 21,841 synsets in WordNet as the visual concepts, and has by far collected 14,197,122 images in total. One limitation of the manually selected concepts is that their visual detectors often fail to capture the complexity of the visual world, and cannot adapt to different domains. For example, people may be interested in detecting birthday cakes when they try to identify a birthday party, but this concept is not present in ImageNet.

To address this problem, we propose to discover the visual concepts automatically by joint use of parallel text and visual corpora. The text data in parallel corpora offers a rich set of terms humans use to describe visual entities, while visual data has the potential to help computer organize the terms into visual concepts. To be useful, we argue that the visual concepts should have the following properties:

- **Discriminative**: a visual concept must refer to visually discriminative entities that can be learned by available computer vision algorithms.
- **Compact**: different terms describing the same set of visual entities should be merged into a single concept.

Our proposed visual concept discovery (VCD) framework first extracts unigrams and dependencies from the text data. It then computes the visual discriminative power of these terms using their associated images and filters out the terms with low cross-validated average precision. The remaining terms may be merged together if they correspond to very similar visual entities. To achieve this, we use semantic similarity and visual similarity scores, and cluster terms based on these similarities. The final output of VCD is a concept vocabulary, where each concept consists a set of terms and has a set of associated images. The pipeline of our approach is illustrated in Figure 1.

We work with the Flickr 8k data set to discover visual concepts; it consists of 8,000 images downloaded from the Flickr website. Each image was annotated by 5 Amazon Mechanical Turk (AMT) workers to describe its content. We design a concept based pipeline for bidirectional image and sentence retrieval task [17] to automatically evaluate the quality of the discovered concepts. We also conduct a human evaluation on a free-form image tagging task using visual concepts. Evaluation results show that the discovered concepts outperform manually selected concepts significantly.

Our key contributions include:

- We show that manually selected concepts often fail to
A black dog and a spotted dog are fighting.
A black dog and a tri-colored dog playing with each other on the road.
Two dogs of different breeds looking at each other on the road.
Two dogs on pavement moving toward each other.

A cyclist is riding a bicycle on a curved road up a hill.
A man on a mountain bike is pedaling up a hill, while cows graze on a hill nearby.
The biker is riding around a curve in the road.

A baseball player attempts to catch a ball while another runs towards the base.

The complexity of and to evolve with the visual world;
- We propose the VCD framework, which automatically generates discriminative and compact visual vocabularies from parallel corpora;
- We demonstrate qualitatively and quantitatively that the discovered concepts outperform several large sets of manually selected concepts significantly. They also perform competitively in the image sentence retrieval task against state-of-the-art embedding based approaches.

2. Related Work

Applications of visual concepts. Visual concepts have been widely used in visual recognition tasks [26, 36, 45]. For example, [11] addresses the problem of describing objects with pre-defined attributes. Sadeghi et al. [37] propose to recognize complex visual composites by defining visual phrases. For video analysis, people commonly use predefined pools of concepts (e.g. blowing candle, cutting cake) to help classify and describe high-level activities or events (e.g. birthday party) [40]. However, their concept vocabularies are usually manually selected.

Concept naming and accuracy-specificity trade-off. Visual concepts can be categorized [34] and organized as a hierarchy where the leaves are the most specific and the root is the most general. For example, ImageNet concepts [6] are organized following the rule-based WordNet [30] hierarchy. Similar structure also exists for actions [9]. Since concept classification is not always reliable, Deng et al. [7] propose a method to allow accuracy-specificity trade-off of object concepts on WordNet. As WordNet synsets do not always correspond to how people name the concepts, Ordonez et al. [31] study the problem of entry-level category prediction by collecting natural categories from humans.

Concept learning from web data. Our research is closely related to the recent work on visual data collection from web images [42, 3, 8, 14] or weakly annotated videos [2]. Their goal is to collect training images from the Internet with minimum human supervision, but for predefined concepts. In particular, NEIL [3] starts with a few exemplar images per concept, and iteratively refines its concept detectors using image search results. LEVAN [8] explores the sub-categories of a given concept by mining bigrams from large text corpora and using the bigrams to retrieve training images from image search engines. Recently, Zhou et al. [44] use noisily tagged Flickr images to train

Figure 1. Overview of the concept discovery framework. Given a parallel corpus of images and their descriptions, we first extract unigrams and dependency bigrams from the text data. These terms are filtered with the cross validation average precision (AP) trained on their associated images. The remaining terms are grouped into concept clusters based on both visual and semantic similarity.
visual concepts. The selected dependent types are likely to correspond to visual concepts. The selected dependent types are likely to correspond to visual concepts. The selected dependent types are likely to correspond to visual concepts. The selected dependent types are likely to correspond to visual concepts.

Sentence generation and retrieval for images. Image descriptions can be generated by detection or retrieval. The detection based approach usually defines a set of visual concepts (e.g. objects, actions and scenes), learns concept detectors and use the top detected concepts to generate sentences. The sentences can be generated using templates [16, 41, 23] or language models [33, 25]. The performance of detection is often limited by missing concepts and inaccurate concept detectors. Retrieval-based sentence generation [32, 24, 43] works by retrieving sentences or sentence components from an existing pool of sentence and image pairs, and use them for description. The retrieval criteria is usually based on the visual similarity of image features. To allow bidirectional retrieval of sentences and images, several work [17, 15, 12] embed image and text raw features into a common latent space using methods like Kernel Canonical Component Analysis [1]. There is also a trend to embed sentences with recurrent neural networks (RNN) [39, 28, 24, 21], which achieves the state-of-the-art performance in sentence retrieval and generation tasks.

3. Visual Concept Discovery Pipeline

This section describes the VCD pipeline. Given a parallel corpus with images and their text descriptions, we first mine the text data to select candidate concepts. Due to the diversity of both visual world and human language, the pool of candidate concepts is large. We use visual data to filter the terms which are not visually discriminative, and then group the remaining terms into compact concept clusters.

3.1. Concept Mining From Sentences

To collect the candidate concepts, we use unigrams as well as the grammatical relations called dependencies [5]. Unlike the syntax tree based representation of sentences, dependencies operate directly on pairs of words. Consider a simple sentence “a little boy is riding a white horse”, white horse and little boy belong to the adjective modifier (amod) dependency, and ride horse belongs to the direct object (dobj) dependency. As the number of dependency types is large, we manually select a subset of 9 types which are likely to correspond to visual concepts. The selected dependency types are: acomp, agent, amod, dobj, iobj, nsubj, nsubjpass, prt and vmod.

The concept mining process proceeds as follows: we first parse the sentences in the parallel corpus with the Stanford CoreNLP parser [5], and collect the terms with the interesting dependency types. We also select unigrams which are annotated as noun, verb, adjective and adverb by a part-of-speech tagger. We use the lemmatized form of the selected unigrams and phrases such that nouns in singular and plural forms and verbs in different tenses are grouped together. After parsing the whole corpus, we remove the terms which occur fewer than $k$ times.

3.2. Concept Filtering and Clustering

The unigrams and dependencies selected from text data contain terms which may not have concrete visual patterns or may not be easy to learn with visual features. The images in the parallel corpora are helpful to filter out these terms. We represent images using feature activations from pre-trained deep convolutional neural networks (CNN), they are image-level holistic features.

Since the number of terms mined from text data is large, the concept filtering algorithm needs to be efficient. For the images associated with a certain term, we do a 2-fold cross validation with a linear SVM, using randomly sampled negative training data. We compute average precision (AP) on cross-validated results, and remove the terms with AP lower than a threshold. Some of the preserved and filtered terms are listed in Table 1.

| Preserved terms          | Filtered terms          |
|-------------------------|-------------------------|
| play tennis, play basketball | play                     |
| bench, kayak             | red bench, blue kayak    |
| sheer, tri-colored       | real, Mexican            |
| biker, dog               | cigar, chess             |

Table 1. Preserved and filtered terms from Flickr 8k data set. A term might be filtered if it’s abstract (first row), too detailed (second row) or not visually discriminative (third row). Sometimes our algorithm may filter out visual entities which are difficult to recognize (final row).
Here the outputs of $f_{t_1}$ are normalized to $[0, 1]$ by a Sigmoid function. We take the minimum of $\hat{S}_w(t_1, t_2)$ and $\hat{S}_v(t_2, t_1)$ to make it a symmetric similarity measurement.

The intuition behind this similarity measurement is that visual instances associated with a term are more likely to get high scores from the classifiers of other visually similar terms.

**Semantic similarity:** We also measure the similarity of two terms in the semantic space, which are computed by data-driven word embeddings. In particular, we train a skip-gram model [29] using the English dump of Wikipedia. The basic idea of skip-gram model is to fit the word embeddings such that the words in corpus can predict their context with high probability. Semantically similar words lie close to each other in the embedded space.

Word embedding algorithm assigns a $D$-dimension vector for each word in the vocabulary. For dependencies, we take the average of the word vectors from each word of the dependency, and $L2$-normalize the averaged vector. The semantic similarity $S_w(t_1, t_2)$ of two candidate concept terms $t_1$ and $t_2$ is defined as the cosine similarity of their word embeddings.

**Concept clustering:** Denote the visual similarity matrix as $S_v$ and the semantic similarity matrix as $S_w$, we compute the overall similarity matrix by

$$S = S_v^\lambda \cdot S_w^{1-\lambda}$$

where $\cdot$ is element-wise matrix multiplication and $\lambda \in [0, 1]$ is a parameter controlling the weight assigned to visual similarity.

We then use spectral clustering to cluster the candidate concept terms into $K$ concept groups. It is a natural choice when similarity matrix is available. We use the algorithm implemented in the Python SKLearn toolkit, fix the eigen solver to arpack and assign the labels with K-means.

After the clustering stage, each concept is represented as a set of terms, as well as their associated visual instances. One can use the associated visual instances to train concept detectors with SVM or neural networks.

| Type     | Concept terms                                      |
|----------|---------------------------------------------------|
| Object   | {jersey, red jersey, orange jersey}              |
| Activity | {dribble, player dribble, dribble ball}           |
| Attribute| {mountainous, hilly}                              |
| Scene    | {blue water, clear water, green water}            |
| Mixed    | {swimming, diving, pool, blue pool}               |
| Mixed    | {ride bull, rodeo, buck, bull}                    |

Table 2. Concepts discovered by our framework from Flickr 8k data set.

| $\lambda$ | Concept terms                                      |
|-----------|---------------------------------------------------|
| 0         | {wedding, church}, {skyscraper, tall building}    |
| 1         | {skyscraper, church}, {wedding, birthday}         |
| 0.3       | {wedding, bridal party}, {church}, {skyscraper}   |

Table 3. Different $\lambda$ affects the term groupings in the discovered concepts. Total concept number is fixed to 1,200.

### 3.3. Discussion

Table 2 shows some of the concepts discovered by our framework. It can automatically generate concepts related to objects, attributes, scenes and activities, and identify the different terms associated with each concept. We observe that sometimes a more general term (jersey) is merged with a more specific term (red jersey) due to high visual similarity.

We also find that there are some mixed type concepts of objects, activities and scenes. For example, swimming and pool belongs to the same concept, possibly due to their high co-occurrence rate. One extreme case is that German and German Shepherd are grouped together as the two words always occur together in the training data. We believe the problem can be mitigated by using a larger parallel corpus.

Table 3 shows different concept clusters when semantic similarity is ignored ($\lambda = 0$), dominant ($\lambda = 1$) and combined with visual similarity. As expected, when $\lambda$ is small, terms that look similar or often co-occur in images tend to be grouped together. As our semantic similarity is based on word co-occurrence, ignoring visual similarity may lead to sub-optimal concept clusters such as wedding and birthday.

### 4. Concept Based Image and Sentence Retrieval

Consider a set of images, each of which has a few ground truth sentence annotations, the goal of bidirectional retrieval is to learn a ranking function from image to sentence and vice versa, such that the ground truth entries rank at the top of the retrieved list. Many previous methods approach the task by learning embeddings from raw feature space [17, 20, 15].

We propose an alternative approach to the embedding based methods which uses concept space directly. Let’s start with the sentence to image direction. With the discovered concepts, this problem can be approached by two steps: first, identify the concepts from the sentences; second, select the images with highest responses for those concepts. Suppose we take the sum of the concept responses, this is equivalent to projecting the sentence into the same concept-based space as images, and measuring the image sentence similarity by an inner product. This formulation allows us to use the same similarity function for image to sentence and sentence to image retrieval.
Sentence mapping: Mapping a sentence to the concept space is straightforward. We run the same parser as used in concept mining to collect terms. Remember that each concept is represented as a set of terms: denote the term set for the incoming sentence as $T = \{t_1, t_2, \ldots, t_N\}$, and the term set for concept $i$ as $C_i = \{c^i_1, c^i_2, \ldots, c^i_M\}$, we have the sentence’s response for $C_i$ as

$$
\phi_i(T) = \max_{t \in T, c \in C_i} \delta(t, c)
$$

Here $\delta(t, c)$ is a function that measures the similarity between $t$ and $c$. We set $\delta(t, c) = 1$ if the cosine similarity of $t$ and $c$’s word embedding is greater than a certain threshold, and 0 otherwise. In practice we set the threshold to 0.95.

There are some common concepts which occur in most of the sentences (e.g. a person); to down-weight these common concepts, we normalize the scores with term frequency-inverse document frequency (tf-idf), learned from the training text corpus.

Image mapping: To measure the response of an image to a certain concept, we need to collect its positive and negative examples. For concepts discovered from parallel corpora, we have their associated images. The set of training images can be augmented with existing image data sets or by manual annotation.

Assume that training images are ready and concept classifiers have been trained, we then compute the continuous classifier scores for an image over all concepts, and normalize each of them to be $[-1, 1]$. The normalization step is important as using non-negative confidence scores biases the system towards longer sentences.

Since image and text data are mapped into a common concept space, the performance of bidirectional retrieval depends on: (1) whether the concept vocabulary covers the terms and visual entities used in query data; (2) whether concept detectors are powerful enough to extract useful information from visual data. It is thus useful to evaluate the quality of discovered concepts against existing concept vocabularies and their concept detectors.

5. Evaluation

In this section, we first evaluate our proposed concept discovery pipeline based on the bidirectional sentence image retrieval task. We use the discovered concepts to generate concept-based image descriptions, and report human evaluation results.

5.1. Bidirectional Sentence Image Retrieval

Data: We use 6,000 images from the Flickr 8k [17] data set for training, 1,000 images for validation and another 1,000 for testing. We use all 5 sentences per image for both training and testing. Flickr 30k [43] is an extended version of Flickr 8k. We select 29,000 images (no overlap to the testing images) to study whether more training data yields better concept detectors. We also report results when the visual concept discovery, concept detector training and evaluation are all conducted on Flickr 30k. For this purpose, we use the standard setting [19, 21] where 29,000 images are used for training, 1,000 images for validation and 1,000 images for testing. Again, each image comes with 5 sentences. Finally, we randomly select 1,000 images from the lately released Microsoft COCO [27] data set to study if the discovered concept vocabulary and associated classifiers generalize to another data set.

Evaluation metric: Recall@k is used for evaluation. It computes the percentage of ground truth entries ranked in the top $k$ retrieved results, over all queries. We also report median rank of the first retrieved ground truth entries.

Image representation and classifier training: Similar to [15, 21], we extracted CNN activations as image-level features; such features have shown state-of-the-art performance in recent object recognition results [22, 13]. We adapted the CNN implementation provided by Caffe [18], and used the 19-layer network architecture and parameters from Oxford [38]. The feature activations from the network’s first fully-connected layer fc6 were used as image representations, each of which has 4,096 dimensions.

To train concept classifiers, we normalized the feature activations with $L2$-norm. We randomly sampled 1,000 images as negative data. We used the linear SVM [10] in the concept discovery stage for its faster running time, and $\chi^2$ kernel SVM to train final concept classifiers as it is a natural choice for histogram-like features and provides higher performance than linear SVM.

Comparison against embedding-based approaches: We first compare the performance of our concept-based pipeline against embedding based approaches. We set the parameters of our system using the validation set. For concept discovery, we kept all terms with at least 5 occurrences in the training sentences, this gave us an initial list of 5,309 terms. We filtered all terms with average precision lower than 0.15, which preserved 2,877 terms. We set $\lambda$ to be 0.6 and number of concepts to be 1,200.

Several recent embedding based approaches [20, 39, 19, 28, 4, 21] are included for comparison. Most of these approaches use CNN-based image representations (in particular, [21] uses the same Oxford architecture), and embed sentences with recurrent neural network (RNN) or its variations. We make sure that the experiment setup and data partitioning for all systems are the same, and report numbers in the original papers if available.

Table 4 lists the evaluation performance for all systems. We can see that the concept based framework achieves similar or better performance against the state-of-the-art embedding based systems. This confirms the framework is a valid pipeline for bidirectional image and sentence retrieval task.
Enhancing concept classifiers with more data: The concept classifiers we trained for previous experiment only used training images from Flickr 8k data set. To check if the discovered concepts can benefit from additional training data, we collect the images associated with the discovered concepts from Flickr 30k data set. Since Flickr 30k contains images which overlap with the validation and testing partitions of Flickr 8k data set, we removed those images and used around 29,000 images for training.

In the last row of Table 4, we list the results of the concept based approach using Flickr 30k training data. We can see that there is a significant improvement in every metric. Since the only difference is the use of additional training data, the results indicate that the individual concept classifiers benefit from extra training data. It is worth noting that while additional data may also be helpful for embedding based approaches, it has to be in the form of image and sentence pairs. Such annotation tends to be more expensive and time consuming to obtain than concept annotation.

Evaluation on Flickr 30k dataset: Evaluation on Flickr 30k follows the same strategy as on Flickr 8k, where parameters were set using validation data. We kept 9,742 terms which have at least 5 occurrences in the training sentences. We then filtered all terms with average precision lower than 0.15, which preserved 4,158 terms. We set λ to be 0.4 and number of concepts to be 1,600. Table 5 shows that our method achieves comparable or better performance than other embedding based approaches.

Concept transfer to other data sets: It is important to investigate whether the discovered concepts are generalizable. For this purpose, we randomly selected 1,000 images and their associated 5,000 text descriptions from the validation partition of Microsoft COCO data set [27].

We used the concepts discovered and trained from Flickr 8k data set, and compared with several existing concept vocabularies:

**ImageNet 1k** [35] is a subset of ImageNet data set, with 1,000 categories used in ILSVRC 2014 evaluation. The classifiers were trained using the same Oxford CNN architecture used for feature extraction.

**LEVAN** [8] selected 305 concepts manually, and explored Google Ngram data to collect 113,983 sub-concepts. They collected Internet images and trained detectors with Deformable Part Model (DPM). We used the learned models provided by the authors.

**NEIL** [3] has 2,702 manually selected concepts, each of which was trained with DPM using weakly supervised images from search engines. We also used the models released by the authors.

Among the three baselines above, ImageNet 1k relies on the same set of CNN-based features as our discovered concepts. To further investigate the effect of concept selection, we took the concept lists provided by the authors of LEVAN and NEIL, and re-trained their concept detectors using our proposed pipeline. To achieve this, we selected training images associated with the concepts from Flickr 8k dataset, and learned concept detectors using the same CNN feature extractors and classifier training strategies as our proposed pipeline.

Table 6 lists the performance of using different vocabularies. We can see that the discovered concepts clearly outperform manually selected vocabularies, but the cross-dataset performance is lower than same-dataset performance. We found that COCO uses many visual concepts discovered in Flickr 8k, though some are missing (e.g. giraffes). Compared with the concepts discovered by Flickr 8k, the three manually selected vocabularies lack many terms used in the COCO data set to describe the visual entities. This inevitably hurts their performance in the retrieval task. The performance of NEIL and LEVAN is worse...
than ImageNet 1k, which might be explained by the weakly
Internet images they used to train concept detectors. Al-
though re-training from Flickr 8k using deep features helps
improve retrieval performance of NEIL and LEVAN, our
system still outperforms the two by large margins.

Impact of concept discovery parameters: Figure 2 and
Figure 3 shows the impact of visual similarity weight $\lambda$ and
the total number of concepts on the retrieval performance.
To save space, we only display results of recall@5 for the
sentence retrieval direction.

We can see from the figures that both visual and semantic similarities are important for concept clustering, this is
particular true when the concepts trained from Flickr 8k
were applied to COCO. Increasing the number of concepts
helps at the beginning, as many visually discriminative concepts are grouped together when the number of concepts is
small. However, as the number increases, the improvement
becomes flat, and even hurts the concepts’ ability to gener-
alize.

5.2. Human Evaluation of Image Tagging

We also evaluated the quality of the discovered concepts
on the image tagging task whose goal is to generate tags to describe the content of images. Compared with sentence
retrieval, the image tagging task has a higher degree of free-
dom as the combination of tags is not limited by the existing
sentences in the pool.

Evaluation setup: We used the concept classifiers to
generate image tags. For each image, we selected the top
concepts with highest classifier scores. Since a con-
cept may have more than one text terms, we picked up to
two text terms randomly for display.

For evaluation, we asked 15 human evaluators to com-
pare two sets of tags generated by different concept vocab-
ularies. The evaluators were asked to select which set of
tags better describes the image based on the accuracy of the
generated tags and the coverage of visual entities in the
image, or whether the two sets of tags are equally good or
bad. The final label per image was combined using major-
ity vote. On average, 85% of the evaluators agreed on their
votes for specific images.

We compared the concepts discovered from Flickr 8k and
the manually selected ImageNet 1k concept vocabulary.
The classifiers for the discovered concepts were trained us-
ing the 6,000 images from Flickr 8k. We did not compare
the discovered concepts against NEIL and LEVAN as they
performed very poorly in the retrieval task. To test how the
concepts generalize to a different data set, we used the same
1,000 images from the COCO data set as used in retrieval

Table 6. Retrieval evaluation for different concept vocabularies on COCO data set.

| Vocabulary                  | Image to sentence |       |       |       | Sentence to image |       |       |       |
|-----------------------------|-------------------|-------|-------|-------|-------------------|-------|-------|-------|
|                             | R@1   | R@5   | R@10  | Median rank | R@1  | R@5   | R@10  | Median rank |
| ImageNet 1k [35]            | 2.5    | 6.7    | 9.7    | 714  | 1.6  | 5.0    | 8.5    | 315  |
| LEVAN [8]                   | 0.0    | 0.4    | 1.2    | 1348 | 0.2  | 1.1    | 1.7    | 443  |
| NEIL [3]                    | 0.1    | 0.7    | 1.1    | 1103 | 0.2  | 0.9    | 2.0    | 446  |
| LEVAN [8] (trained on Flickr 8k) | 1.2  | 5.7    | 9.5    | 360  | 2.6  | 9.1    | 14.7   | 113  |
| NEIL [3] (trained on Flickr 8k) | 1.4  | 5.7    | 8.9    | 278  | 3.7  | 11.3   | 18.3   | 92   |
| Flickr 8k Concepts (ours)   | 10.4   | 29.3   | 40.0   | 17   | 9.8  | 27.5   | 39.0   | 17   |

Figure 2. Impact of $\lambda$ when testing on Flickr 8k data set (blue) and COCO data set (red). Recall@5 for sentence retrieval is used.

Figure 3. Impact of total number of concepts when testing on Flickr 8k data set (blue) and COCO data set (red). Recall@5 for sentence retrieval is used.
| Better | Worse | Same |
|--------|-------|------|
| 64.1%  | 22.9% | 12.9%|

Table 7. Percentage of images where tags generated by the discovered concepts are better, worse or the same compared with ImageNet 1k.

Figure 4. Tags generated using ImageNet 1k concepts (blue) and the discovered concepts (green). Tags preferred by evaluators are marked in red blocks.

Result analysis: Table 7 shows the evaluators’ preference on the image tags generated by the discovered concepts and ImageNet 1k. We can see that the discovered concepts generated better tags for 64.1% of the images. This agrees with the trend observed in the bidirectional retrieval task.

As shown in Figure 4, tags generated by ImageNet 1k has the following problems which might cause evaluators to label them as worse: first, many of the visual entities do not have corresponding concepts in the vocabulary; second, ImageNet 1k has many fine-grained concepts (e.g. different species of dogs), while more general terms might be preferred by evaluators. On the other hand, the discovered concepts are able to reflect how human name the visual entities, and have a higher concept coverage. However, due to the number of training examples is relatively limited, sometimes the response of different concept classifiers are correlated (e.g. bed and sit down).

6. Conclusion

This paper studies the problem of automatic concept discovery from parallel corpora. We propose a concept filtering and clustering algorithm using both text and visual information. Automatic evaluation using bidirectional image and text retrieval and human evaluation of image tagging task show that the discovered concepts achieve state-of-the-art performance, and outperform several large manually selected concept vocabularies significantly. A natural future direction is to train concept detectors for the discovered concepts using web images.

Acknowledgement: We thank Kevin Knight for helpful discussions. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. JMLR, 2002.
[2] C.-Y. Chen and K. Grauman. Watching unlabeled video helps learn new human actions from very few labeled snapshots. In CVPR, 2013.
[3] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In ICCV, 2013.
[4] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014.
[5] M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In LREC, 2006.
[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009.
