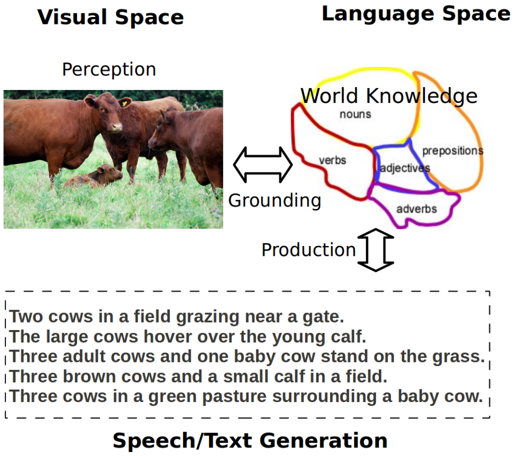
Corpus-Guided Sentence Generation of Natural Images

Yezhou Yang[†] and **Ching Lik Teo**[†] and **Hal Daumé III** and **Yiannis Aloimonos**
University of Maryland Institute for Advanced Computer Studies
College Park, Maryland 20742, USA
{yzyang, cteo, hal, yiannis}@umiacs.umd.edu

Abstract

We propose a sentence generation strategy that describes images by predicting the most likely nouns, verbs, scenes and prepositions that make up the core sentence structure. The input are initial noisy estimates of the objects and scenes detected in the image using state of the art trained detectors. As predicting actions from still images directly is unreliable, we use a language model trained from the English Gigaword corpus to obtain their estimates; together with probabilities of co-located nouns, scenes and prepositions. We use these estimates as parameters on a HMM that models the sentence generation process, with hidden nodes as sentence components and image detections as the emissions. Experimental results show that our strategy of combining vision and language produces readable and descriptive sentences compared to naive strategies that use vision alone.



Two cows in a field grazing near a gate.
The large cows hover over the young calf.
Three adult cows and one baby cow stand on the grass.
Three brown cows and a small calf in a field.
Three cows in a green pasture surrounding a baby cow.

Figure 1: The processes involved for describing a scene.

1 Introduction

What happens when you see a picture? The most natural thing would be to *describe* it using *words*: using speech or text. This description of an image is the output of an extremely complex process that involves: 1) perception in the Visual space, 2) grounding to World Knowledge in the Language Space and 3) speech/text production (see Fig. 1). Each of these components are challenging in their own right and are still considered open problems in the vision and linguistics fields. In this paper, we introduce a computational framework that attempts to integrate these components together. Our hypothesis is based on the assumption that natural images accurately reflect common everyday scenarios which are captured in language. For example, knowing that boats usually occur over `water` will enable us to constrain the possible scenes a boat can occur and exclude highly unlikely ones – `street, highway`. It also enables us to predict likely actions (Verbs) given the current object detections in the image: detecting a dog with a person will likely induce `walk` rather than `swim, jump, fly`. Key to our approach is the use of a large generic corpus such as the English Gigaword [Graff, 2003] as the *semantic grounding* to predict and correct the initial and often noisy visual detections of an image to produce a reasonable sentence that succinctly describes the image.

In order to get an idea of the difficulty of this task, it is important to first define what makes up

[†] indicates equal contribution.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 444–454,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

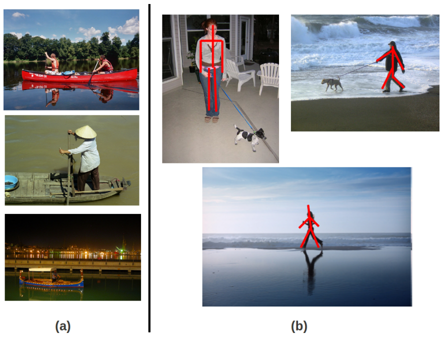


Figure 2: Illustration of various perceptual challenges for sentence generation for images. (a) Different images with semantically the same content. (b) Pose relates ambiguously to actions in real images. See text for details.

a description of an image. Based on our observations of annotated image data (see Fig. 4), a descriptive sentence for an image must contain at minimum: 1) the important *objects* (Nouns) that participate in the image, 2) Some description of the *actions* (Verbs) associated with these objects, 3) the *scene* where this image was taken and 4) the *preposition* that relates the objects to the scene. That is, a quadruplet of $\mathcal{T} = \{n, v, s, p\}$ (Noun-Verb-Scene-Preposition) that represents the core sentence structure. Generating a sentence from this quadruplet is obviously a simplification from state of the art generation work, but as we will show in the experimental results (sec. 4), it is sufficient to describe images. The key challenge is that detecting objects, actions and scenes *directly* from images is often noisy and unreliable. We illustrate this using example images from the Pascal-Visual Object Classes (VOC) 2008 challenge [Everingham et al., 2008]. First, Fig. 2(a) shows the *variability* of images in their raw image representations: pixels, edges and local features. This makes it difficult for state of the art object detectors [Felzenszwalb et al., 2010; Schwartz et al., 2009] to reliably detect important objects in the scene: boat, humans and water – average precision scores reported in [Felzenszwalb et al., 2010] manages around 42% for humans and only 11% for boat over a dataset of almost 5000 images in 20 object categories. Yet, these images are *semantically* similar in terms of their high level description. Second, cognitive studies [Urgesi et al., 2006; Kourtzi, 2004] have proposed that inferring the action from static images (known as an "implied action") is often achieved by detecting the *pose* of humans in the image: the position of the limbs with respect to one another, under the assumption that a unique pose occurs for a unique action. Clearly, this assumption is weak as 1) similar actions may be represented by different poses due to the inherent dynamic nature of the action itself: e.g. walking a dog and 2) different actions may have the same pose: e.g. walking a dog versus running (Fig. 2(b)). The missing component here is whether the key object (dog) under interaction is considered. Recent works [Yao and Fei-Fei, 2010; Yang et al., 2010] that used poses for recognition of actions achieved 70% and 61% accuracy respectively under extremely limited testing conditions with only 5-6 action classes each. Finally, state of the art scene detectors [Oliva and Torralba, 2001; Torralba et al., 2003] need to have enough representative training examples of scenes from pre-defined scene classes for a classification to be successful – with a reported average precision of 83.7% tested over a dataset of 2600 images.

Addressing all these visual challenges is clearly a formidable task which is beyond the scope of this paper. Our focus instead is to show that with the addition of *language* to ground the noisy initial visual detections, we are able to improve the quality of the generated sentence as a faithful description of the image. In particular, we show that it is possible to avoid predicting actions directly from images – which is still unreliable – and to use the corpus instead to guide our predictions. Our proposed strategy is also *generic*, that is, we make no prior assumptions on the image domain considered. While other works (sec. 2) depend on strong annotations between images and text to ground their predictions (and to remove wrong sentences), we show that a large generic corpus is also able to provide the same grounding over larger domains of images. It represents a relatively new style of learning: distant supervision [Liang et al., 2009; Mann and Mccallum, 2007]. Here, we do not require "labeled" data containing images and captions but only separate data from each side. Another contribution is a computationally feasible way via dynamic programming to determine the most likely quadruplet $\mathcal{T}^* = \{n^*, v^*, s^*, p^*\}$ that describes the image for generating possible sentences.

2 Related Work

Recently, several works from the Computer Vision domain have attempted to use language to aid image scene understanding. [Kojima et al., 2000] used predefined production rules to describe actions in videos. [Berg et al., 2004] processed news captions to discover names associated with faces in the images, and [Jie et al., 2009] extended this work to associate poses detected from images with the verbs in the captions. Both approaches use annotated examples from a limited news caption corpus to learn a joint image-text model so that one can annotate new unknown images with textual information easily. Neither of these works have been tested on complex everyday images where the large variations of objects and poses makes it nearly impossible to learn a more general model. In addition, no attempt was made to generate a descriptive sentence from the learned model. The work of [Farhadi et al., 2010] attempts to "generate" sentences by first learning from a set of human annotated examples, and producing the *same* sentence if both images and sentence share common properties in terms of their triplets: (Nouns-Verbs-Scenes). No attempt was made to generate *novel* sentences from images beyond what has been annotated by humans. [Yao et al., 2010] has recently introduced a framework for parsing images/videos to textual description that requires significant annotated data, a requirement that our proposed approach avoids.

Natural language generation (NLG) is a long-standing problem. Classic approaches [Traum et al., 2003] are based on three steps: selection, planning and realization. A common challenge in generation problems is the question of: what is the input? Recently, approaches for generation have focused on formal specification inputs, such as the output of theorem provers [McKeown, 2009] or databases [Golland et al., 2010]. Most of the effort in those approaches has focused on selection and realization. We address a tangential problem that has not received much attention in the generation literature: how to deal with *noisy inputs*. In our case, the inputs themselves are often uncertain (due to misrecognitions by object/scene detectors) and the content selection and realization needs to take this uncertainty into account.

3 Our Approach

Our approach is summarized in Fig. 3. The input is a test image where we detect objects and scenes using trained detection algorithms [Felzenszwalb et al., 2010; Torralba et al., 2003]. To keep the framework computationally tractable, we limit the elements of the quadruplet (Nouns-Verbs-Scenes-Prepositions) to come from a finite set of objects $\mathcal{N}$, actions $\mathcal{V}$, scenes $\mathcal{S}$ and prepositions $\mathcal{P}$ classes that are commonly encountered. They are summarized in Table. 1. In addition, the sentence that is generated for each image is limited to at most two objects occurring in a unique scene.

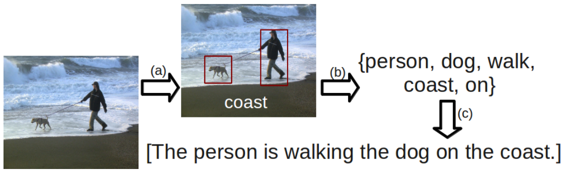
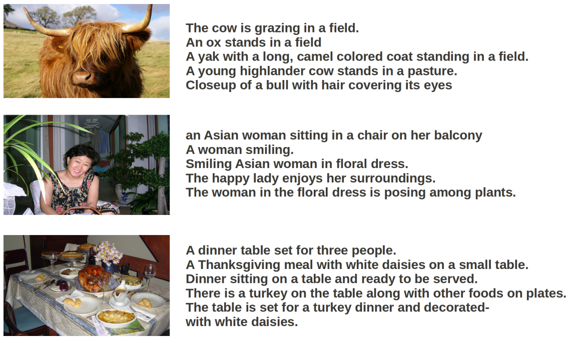

(a)
coast
(b)
{person, dog, walk, coast, on}
(c)
[The person is walking the dog on the coast.]

Figure 3: Overview of our approach. (a) Detect objects and scenes from input image. (b) Estimate optimal sentence structure quadruplet $\mathcal{T}^*$. (c) Generating a sentence from $\mathcal{T}^*$.

Denoting the current test image as I, the initial visual processing first detects objects $n \in \mathcal{N}$ and scenes $s \in \mathcal{S}$ using these detectors to compute $P_r(n|I)$ and $P_r(s|I)$, the probabilities that object n and scene s exist under I. From the observation that an action can often be predicted by its key objects, $N_k = \{n_1, n_2, \cdots, n_i\}, n_i \in \mathcal{N}$ that participate in the action, we use a trained Language model L_m to estimate $P_r(v|N_k)$. L_m is also used to compute $P_r(s|n, v)$, the predicted scene using the corpus given the object and verb; and $P_r(p|s)$, the predicted preposition given the scene. This process is repeated over all n, v, s, p where we used a modified HMM inference scheme to determine the most likely quadruplet: $\mathcal{T}^* = \{n^*, v^*, s^*, p^*\}$ that makes up the core sentence structure. Using the contents and structure of $\mathcal{T}^*$, an appropriate sentence is then generated that describes the image. In the following sections, we first introduce the image dataset used for testing followed by details of how these components are derived.

Objects $n \in \mathcal{N}$	Actions $v \in \mathcal{V}$	Scenes $s \in \mathcal{S}$	Preps $p \in \mathcal{P}$
'aeroplane' 'bicycle' 'bird'	'sit' 'stand' 'park'	'airport'	'in' 'at' 'above'
'boat' 'bottle' 'bus' 'car'	'ride' 'hold' 'wear'	'field'	'around' 'behind'
'cat' 'chair' 'cow' 'table'	'pose' 'fly' 'lie' 'lay'	'highway'	'below' 'beside'
'dog' 'horse', 'motorbike'	'smile' 'live' 'walk'	'lake' 'room'	'between'
'person' 'pottedplant'	'graze' 'drive' 'play'	'sky' 'street'	'before' 'to'
'sheep' 'sofa' 'train'	'eat' 'cover' 'train'	'track'	'under' 'on'
'tvmonitor'	'close' ...		

Table 1: The set of objects, actions (first 20), scenes and preposition classes considered


The cow is grazing in a field.
An ox stands in a field
A yak with a long, camel colored coat standing in a field.
A young highlander cow stands in a pasture.
Closeup of a bull with hair covering its eyes

an Asian woman sitting in a chair on her balcony
A woman smiling.
Smiling Asian woman in floral dress.
The happy lady enjoys her surroundings.
The woman in the floral dress is posing among plants.

A dinner table set for three people.
A Thanksgiving meal with white daisies on a small table.
Dinner sitting on a table and ready to be served.
There is a turkey on the table along with other foods on plates.
The table is set for a turkey dinner and decorated-
with white daisies.

Figure 4: Samples of images with corresponding annotations from the UIUC scene description dataset.

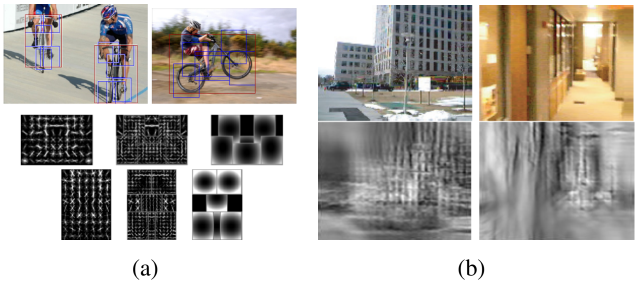


Figure 5: (a) [Top] The part based object detector from [Felzenszwalb et al., 2010]. [Bottom] The graphical model representation of an object, for e.g. a bike. (b) Examples of GIST gradients: (left) an outdoor scene vs (right) an indoor scene [Torralba et al., 2003].

3.1 Image Dataset

We use the *UIUC Pascal Sentence dataset*, first introduced in [Farhadi et al., 2010] and available online[1]. It contains 1000 images taken from a subset of the Pascal-VOC 2008 challenge image dataset and are hand annotated with sentences that describe the image by paid human annotators using Amazon Mechanical Turk. Fig. 4 shows some sample images with their annotations. There are 5 annotations per image, and each annotation is usually short – around 10 words long. We randomly selected 900 images (4500 sentences) as the learning corpus to construct the verb and scene sets, $\{\mathcal{V}, \mathcal{S}\}$ as described in sec. 3.3, and kept the remaining 100 images for testing and evaluation.

3.2 Object and Scene Detections from Images

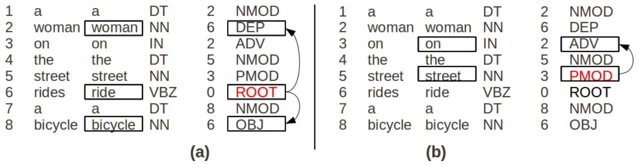
We use the Pascal-VOC 2008 trained object detectors [Felzenszwalb et al., 2008] of 20 common everyday object classes that are defined in $\mathcal{N}$. Each of the detectors are essentially SVM classifiers trained on a large number of the objects' image representations from a large variety of sources. Although 20 classes may seem small, their existence in many natural images (e.g. humans, cars and plants) makes them particularly *important* for our task, since humans tend to describe these common objects as well. As object representations, the part-based descriptor of [Felzenszwalb et al., 2010] is used. This representation decomposes any object, e.g. a cow, into its constituent parts: head, torso, legs, which are shared by other objects in a hierarchical manner. At each level, image gradient orientations are computed. The relationship between each parts is modeled probabilistically using graphical models where parts are the nodes and the edges are the conditional probabilities that relate their spatial compatibility (Fig. 5(a)). For example, in a cow, the probability of finding the torso near the head is higher than finding the legs near the head. This model's intuition lies in the assumption that objects can be deformed but the relative position of each constituent parts should remain the same. We convert the object detection scores to probabilities using Platt's method [Lin et al., 2007] which is numerically more stable to obtain $P_r(n|I)$. The parameters of Platt's method are obtained by estimating the number of positives and negatives from the UIUC annotated dataset, from

[1] http://vision.cs.uiuc.edu/pascal-sentences/

which we determine the appropriate probabilistic threshold, which gives us approximately 50% recall and precision.

For detecting scenes defined in $\mathcal{S}$, we use the GIST-based scene descriptor of [Torralba et al., 2003]. GIST computes the windowed 2D Gabor filter responses of an input image. The responses of Gabor filters (4 scales and 6 orientations) encode the texture gradients that describe the *local* properties of the image. Averaging out these responses over larger spatial regions gives us a set of *global* image properties. These high dimensional responses are then reprojected to a low dimensional space via PCA, where the number of principal components are obtained empirically from training scenes. This representation forms the GIST descriptor of an image (Fig. 5(b)) which is used to train a set of SVM classifiers for each scene class in $\mathcal{S}$. Again, $P_r(s|I)$ is computed from the SVM scores using [Lin et al., 2007]. The set of common scenes defined in $\mathcal{S}$ is learned from the UIUC annotated data (sec. 3.3).

3.3 Corpus-Guided Predictions



1	a	a	DT	2	NMOD
2	woman	woman	NN	6	DEP
3	on	on	IN	2	ADV
4	the	the	DT	5	NMOD
5	street	street	NN	3	PMOD
6	rides	ride	VBZ	0	ROOT
7	a	a	DT	8	NMOD
8	bicycle	bicycle	NN	6	OBJ

(a)

1	a	a	DT	2	NMOD
2	woman	woman	NN	6	DEP
3	on	on	IN	2	ADV
4	the	the	DT	5	NMOD
5	street	street	NN	3	PMOD
6	rides	ride	VBZ	0	ROOT
7	a	a	DT	8	NMOD
8	bicycle	bicycle	NN	6	OBJ

(b)

Figure 6: (a) Selecting the ROOT verb from the dependency parse `ride` reveals its subject `woman` and direct object `bicycle`. (b) Selecting the head noun (PMOD) as the scene `street` reveals ADV as the preposition `on`

Predicting Verbs: The key component of our approach is the trained language model L_m that predicts the most likely verb v, associated with the objects N_k detected in the image. Since it is possible that different verbs may be associated with varying number of object arguments, we limit ourselves to verbs that take on at most *two* objects (or more specifically two noun phrase arguments) as a simplifying assumption: $N_k = \{n_1, n_2\}$ where n_2 can be NULL. That is, n_1 and n_2 are the subject and direct objects associated with $v \in \mathcal{V}$. Using this assumption, we can construct the set of verbs, $\mathcal{V}$. To do this, we use human labeled descriptions of the training images from the UIUC Pascal-VOC dataset (sec. 3.1) as a learning corpus that allows us to determine the appropriate target verb set that is amenable to our problem. We first apply the CLEAR parser [Choi and Palmer, 2010] to obtain a dependency parse of these annotations, which also performs stemming of all the verbs and nouns in the sentence. Next, we process all the parses to select verbs which are marked as ROOT and check the existence of a subject (DEP) and direct object (PMOD, OBJ) that are linked to the ROOT verb (see Fig. 6(a)). Finally, after removing common "stop" verbs such as {`is, are, be`} we rank these verbs in terms of their occurrences and select the top 50 verbs which accounts for 87.5% of the sentences in the UIUC dataset to be in $\mathcal{V}$.

Object class $n \in \mathcal{N}$	Synonyms, $\langle n \rangle$
`bus`	`autobus charabanc double-decker jitney motorbus motorcoach omnibus passenger-vehicle schoolbus trolleybus streetcar ...`
`chair`	`highchair chaise daybed throne rocker armchair wheelchair seat ladder-back lawn-chair fauteuil ...`
`bicycle`	`bike wheel cycle velocipede tandem mountain-bike ...`

Table 2: Samples of synonyms for 3 object classes.

Next, we need to explain how n_1 and n_2 are selected from the 20 object classes defined previously in $\mathcal{N}$. Just as the 20 object classes are defined *visually* over several different kinds of specific objects, we expand n_1 and n_2 in their textual descriptions using *synonyms*. For example, the object class n_1=`aeroplane` should include the synonyms {`plane, jet, fighter jet, aircraft`}, denoted as $\langle n_1 \rangle$. To do this, we expand each object class using their corresponding WordNet synsets up to at most three hyponymns levels. Example synonyms for some of the classes are summarized in Table 2.

We can now compute from the Gigaword corpus [Graff, 2003] the probability that a verb exists given the detected nouns, $P_r(v|n_1, n_2)$. We do this by computing the log-likelihood ratio [Dunning, 1993] , λ_{nvn}, of *trigrams* $(\langle n_1 \rangle, v, \langle n_2 \rangle)$, computed from each sentence in the English Gigaword corpus [Graff, 2003]. This is done by extracting only the words in the corpus that are defined in $\mathcal{N}$ and $\mathcal{V}$ (including their synonyms). This forms a *reduced* corpus sequence from which we obtain our target trigrams. For example, the sentence:

```
the large brown dog chases a small young cat
around the messy room, forcing the cat to run
away towards its owner.
```

will be reduced to the stemmed sequence `dog chase cat cat run owner`[2] from which we obtain the target trigram relationships: {`dog chase cat`}, {`cat run owner`} as these trigrams respect the (n_1, v, n_2) ordering. The log-likelihood ratios, λ_{nvn}, computed for all possible $(\langle n_1 \rangle, v, \langle n_2 \rangle)$ are then normalized to obtain $P_r(v|n_1, n_2)$. An example of ranked λ_{nvn} in Fig. 7(a) shows that λ_{nvn} predicts v that makes sense: with the most likely predictions near the top of the list.

Predicting Scenes: Just as an action is strongly related to the objects that participate in it, a scene can be predicted from the objects and verbs that occur in the image. For example, detecting N_k={`boat, person`} with v={`row`} would have predicted the scene s={`coast`}, since boats usually occur in water regions. To learn this relationship from the corpus, we use the UIUC dataset to discover what are the common scenes that should be included in $\mathcal{S}$. We applied the CLEAR dependency parse [Choi and Palmer, 2010] on the UIUC data and extracted all the head nouns (PMOD) in the PP phrases for this purpose and excluded those nouns with prepositions (marked as ADV) such as {`with, of`} which do not co-occur with scenes in general (see Fig. 6(b)). We then ranked the remaining scenes in terms of their frequency to select the top 8 scenes used in $\mathcal{S}$.

To improve recall and generalization, we expand each of the 8 scene classes using their WordNet synsets $\langle s \rangle$ (up to a max of three hyponymns levels). Similar to the procedure of predicting the verbs described above, we compute the log-likelihood ratio of ordered *bigrams*, $\{n, \langle s \rangle\}$ and $\{v, \langle s \rangle\}$: λ_{ns} and λ_{vs}, by reducing the corpus sentence to the target nouns, verbs and scenes defined in $\mathcal{N}, \mathcal{V}$ and $\mathcal{S}$. The probabilities $P_r(s|n)$ and $P_r(v|n)$ are then obtained by normalizing λ_{ns} and λ_{vs}. Under the assumption that the priors $P_r(n)$ and $P_r(v)$ are independent and applying Bayes rule, we can compute the probability that a scene co-occurs with the object and action, $P_r(s|n, v)$ by:

$$\begin{aligned} P_r(s|n,v) &= \frac{P_r(n,v|s)P_r(s)}{P_r(n,v)} \\ &= \frac{P_r(n|s)P_r(v|s)P_r(s)}{P_r(n)P_r(v)} \\ &\propto P_r(s|n) \times P_r(s|v) \qquad (1) \end{aligned}$$

where the constant of proportionality is justified under the assumption that $P_r(s)$ is equiprobable for all s. (1) is computed for all nouns in N_k. As shown in Fig. 7(b), we are able to predict scenes that colocate with reasonable correctness given the nouns and verbs.

Predicting Prepositions: It is straightforward to predict the appropriate prepositions associated with a given scene. When we construct $\mathcal{S}$ from the UIUC annotated data, we simply collect and rank all the associated prepositions (ADV) in the PP phrase of the dependency parses. We then select the top 12 prepositions used to define $\mathcal{P}$. Using $\mathcal{P}$, we then compute the log-likelihood ratio of ordered *bigrams*, $\{p, \langle s \rangle\}$ for prepositions that co-locate with the scene synonyms over the corpus. Normalizing λ_{ps} yields $P_r(p|s)$, the probability that a preposition co-locates with a scene. Examples of ranked λ_{ps} are shown in Fig. 7(c). Again, we see that reasonable predictions of p can be found.

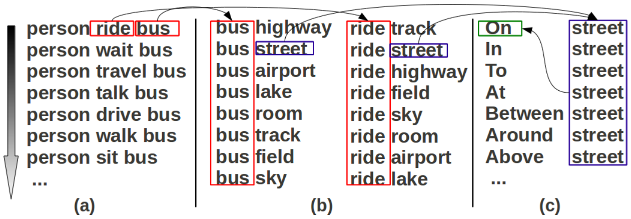


Figure 7: Example of how ranked log-likelihood values (in descending order) suggest a possible $\mathcal{T}$: (a) λ_{nvn} for $n_1 = \texttt{person}, n_2 = \texttt{bus}$ predicts $v = \texttt{ride}$. (b) λ_{ns} and λ_{vs} for $n = \texttt{bus}, v = \texttt{ride}$ then jointly predicts $s = \texttt{street}$ and finally (c) λ_{ps} with $s = \texttt{street}$ predicts $p = \texttt{on}$.

3.4 Determining $\mathcal{T}^*$ using HMM inference

Given the computed conditional probabilities: $P_r(n|I)$ and $P_r(s|I)$ which are observations from an input test image with the parameters of the trained language model, L_m:

[2] stemming is done using [Choi and Palmer, 2010]

$P_r(v|n_1, n_2), P_r(s|n, v), P_r(p|s)$, we seek to find the most likely sentence structure $\mathcal{T}^*$ by:

$$\begin{aligned} \mathcal{T}^* &= \arg\max_{n,v,s,p} P_r(\mathcal{T}|n, v, s, p) \\ &= \arg\max_{n,v,s,p} \{P_r(n_1|I)P_r(n_2|I)P_r(s|I)\times \\ &\qquad P_r(v|n_1, n_2)P_r(s|n, v)P_r(p|s)\} \end{aligned} \quad (2)$$

where the last equality holds by assuming independence between the visual detections and corpus predictions. Obviously a brute force approach to try all possible combinations to maximize eq. (2) will not be feasible due to the large number of possible combinations: $(20 * 21 * 8) * (50 * 20 * 20) * (8 * 20 * 50) * (12 * 8) \approx 5 \times 10^{13}$. A better solution is needed.

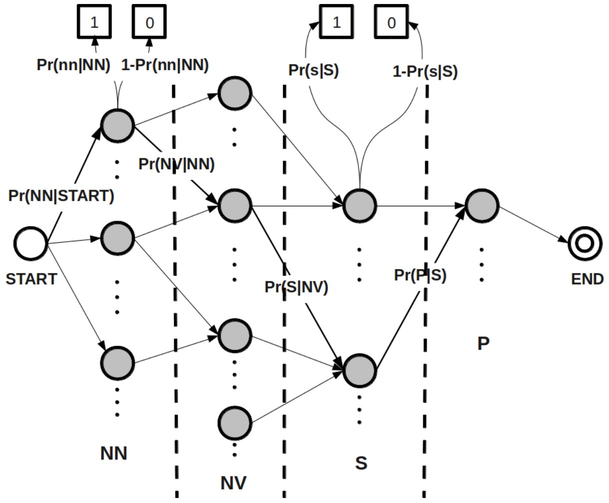


Figure 8: The HMM used for optimizing $\mathcal{T}$. The relevant transition and emission probabilities are also shown. See text for more details.

Our proposed strategy is to pose the optimization of $\mathcal{T}$ as a dynamic programming problem, akin to a Hidden Markov Model (HMM) where the hidden states are related to the (simplified) sentence structure we seek: $\mathcal{T} = \{n_1, n_2, s, v, p\}$, and the emissions are related to the observed detections: $\{n_1, n_2, s\}$ in the image if they exist. To simplify our notations, as we are concerned with object pairs we will write NN as the hidden states for all n_1, n_2 pairs and nn as the corresponding emissions (detections); and all object+verb pairs as hidden states NV. The hidden states are therefore denoted as: $\{\mathtt{NN}, \mathtt{NV}, \mathtt{S}, \mathtt{P}\}$ with values taken from their respective word classes from Table 1. The emission states are $\{\mathtt{nn}, \mathtt{s}\}$ with binary values: 1 if the detections occur or 0 otherwise. The full HMM is summarized in Fig. 8. The rationale for using a HMM is that we can reuse all previous computation of the probabilities at each level to compute the required probabilities at the current level. From START, we assume all object pair detections are equiprobable: $P_r(\mathtt{NN}|\mathtt{START}) = \frac{1}{|\mathcal{N}|*(|\mathcal{N}|+1)}$ where we have added an additional NULL value for objects (at most 1). At each NN, the HMM emits a detection from the image and by independence we have: $P_r(\mathtt{nn}|\mathtt{NN}) = P_r(n_1|I)P_r(n_2|I)$. After NN, the HMM transits to the corresponding verb at state NV with $P_r(\mathtt{NV}|\mathtt{NN}) = P_r(v|n_1, n_2)$ obtained from the corpus statistic[3]. As no action detections are performed on the image, NV has no emissions. The HMM then transits from NV to S with $P_r(\mathtt{S}|\mathtt{NV}) = P_r(s|n, v)$ computed from the corpus which emits the scene detection score from the image: $P_r(\mathtt{s}|\mathtt{S}) = P_r(s|I)$. From S, the HMM transits to P with $P_r(\mathtt{P}|\mathtt{S}) = P_r(p|s)$ before reaching the END state.

Comparing the HMM with eq. (2), one can see that all the corpus and detection probabilities are accounted for in the transition and emission probabilities respectively. Optimizing $\mathcal{T}$ is then equivalent to finding the best (most likely) path through the HMM given the image observations using the Viterbi algorithm which can be done in $O(10^5)$ time which is significantly faster than the naive approach. We show in Fig. 9 (right-upper) examples of the top viterbi paths that produce $\mathcal{T}^*$ for four test images.

Note that the proposed HMM is suitable for generating sentences that contain the core components defined in $\mathcal{T}$ which produces a sentence of the form NP-VP-PP, which we will show in sec. 4 is sufficient for the task of generating sentences for describing images. For more complex sentences with more components: such as adjectives or adverbs, the HMM can be easily extended with similar computations derived from the corpus.

3.5 Sentence Generation

Given the selected sentence structure $\mathcal{T} = \{n_1, n_2, v, s, p\}$, we generate sentences using the

[3] each verb, v, in NV will have 2 entries with the same value, one for each noun.

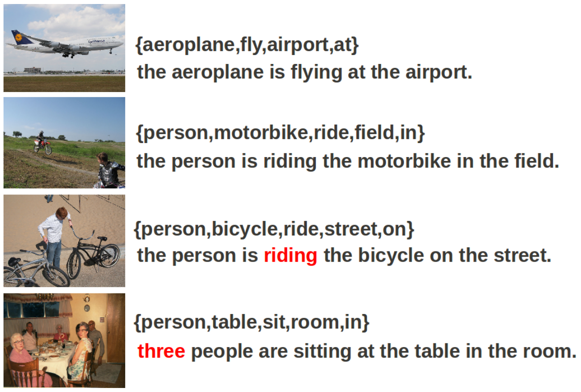

{aeroplane,fly,airport,at}
the aeroplane is flying at the airport.
{person,motorbike,ride,field,in}
the person is riding the motorbike in the field.
{person,bicycle,ride,street,on}
the person is riding the bicycle on the street.
{person,table,sit,room,in}
three people are sitting at the table in the room.

Figure 9: Four test images (left) and results. (Right-upper): Sentence structure $\mathcal{T}^*$ predicted using Viterbi and (Right-lower): Generated sentences. Words marked in red are considered to be incorrect predictions. Complete results are available at `http://www.umiacs.umd.edu/~yzyang/sentence_generateOut.html`.

following strategy for each component:

1) We add in appropriate determiners and cardinals: `the, an, a, CARD`, based on the content of n_1,n_2 and s. For e.g., if $n_1 = n_2$, we will use `CARD=two`, and modify the nouns to be in the plural form. When several possible choices are available, a random choice is made that depends on the object detection scores: `the` is preferred when we are confident of the detections while `an, a` is preferred otherwise.

2) We predict the most likely preposition inserted between the verbs and nouns learned from the Gigaword corpus via $P_r(p|v,n)$ during sentence generation. For example, our method will pick the preposition `at` between verb `sit` and noun `table`.

3) The verb v is converted to a form that agrees with in number with the nouns detected. The present gerund form is preferred such as `eating, drinking, walking` as it conveys that an action is being performed in the image.

4) The sentence structure is therefore of the form: `NP-VP-PP` with variations when only one object or multiple detections of the same objects are detected. A special case is when *no* objects are detected (below the predefined threshold). No verbs can be predicted as well. In this case, we simply generate a sentence that describes the *scene* only: for e.g. `This is a coast, This is a field`. Such sentences account for 20% of the entire UIUC testing dataset which are scored lower in our evaluation metrics (sec. 4.1) since they do not fully *describe* the image content in terms of the objects and actions.

Some examples of sentences generated using this strategy are shown in Fig. 9(right-lower).

4 Experiments

We performed several experiments to evaluate our proposed approach. The different metrics used for evaluation and comparison are also presented, followed by a discussion of the experimental results.

4.1 Sentence Generation Results

Three experiments are performed to evaluate the effectiveness of our approach. As a baseline, we simply generated $\mathcal{T}^*$ *directly* from images without using the corpus. There are two variants of this baseline where we seek to determine if listing *all* objects in the image is crucial for scene description. $\mathcal{T}_{b1}$ is a baseline that uses *all* possible objects and scene detected: $\mathcal{T}_{b1} = \{n_1, n_2, \cdots, n_m, s\}$ and our sentence will be of the form: {`Object 1, object 2 and object 3 are IN the scene.`} and we simply selected `IN` as the only admissible preposition. For the second baseline, $\mathcal{T}_{b2}$, we limit the number of objects to just any two: $\mathcal{T}_{b2} = \{n_1, n_2, s\}$ and the sentence generated will be of the form {`Object 1 and object 2 are IN the scene`}. In the second experiment, we applied the HMM strategy described above but made all transition probabilities *equiprobable*, removing the effects of the corpus, and producing a sentence structure which we denote as $\mathcal{T}^*_{eq}$. The third experiment produces the full $\mathcal{T}^*$ with transition probabilities learned from the corpus. All experiments were performed on the 100 unseen testing images from the UIUC dataset and we used only the most likely (top) sentence generated for all evaluation.

We use two evaluation metrics as a measure of the accuracy of the generated sentences: 1) ROUGE-1 [Lin and Hovy, 2003] precision scores and 2) *Relevance* and *Readability* of the generated sentences. ROUGE-1 is a recall based metric that is commonly used to measure the effectiveness of text summarization. In this work, the short descriptive sentence of an image can be viewed as summarizing the image content and ROUGE-1 is able to capture how well this sentence can describe the image by comparing it with the human annotated ground truth of the UIUC dataset. Due to the short sentences generated, we did not consider other ROUGE metrics (ROUGE-2, ROUGE-SU4) which captures fluency and is not an issue here.

Experiment	R_1,(length)	Relevance	Readability
Baseline 1, $\mathcal{T}^*_{b1}$	0.35,(8.2)	$\mathbf{2.84} \pm 1.40$	3.64 ± 1.20
Baseline 2, $\mathcal{T}^*_{b2}$	0.39,(6.8)	2.14 ± 1.13	3.94 ± 0.91
HMM no corpus, $\mathcal{T}^*_{eq}$	0.42,(6.5)	2.44 ± 1.25	3.88 ± 1.18
Full HMM, $\mathcal{T}^*$	$\mathbf{0.44}$,(6.9)	2.51 ± 1.30	$\mathbf{4.10} \pm 1.03$
Human Annotation	0.68,(10.1)	4.91 ± 0.29	4.77 ± 0.42

Table 3: Sentence generation evaluation results with human gold standard. Human R_1 scores are averaged over the 5 sentences using a leave one out procedure. Values in bold are the top scores.

A main shortcoming of using ROUGE-1 is that the generated sentences are compared only to a finite set of human labeled ground truth which obviously does not capture all possible sentences that one can generate. In other words, ROUGE-1 does not take into account the fact that sentence generation is innately a *creative* process, and a better recall metric will be to ask humans to judge these sentences. The second evaluation metric: Relevance and Readability is therefore proposed as an empirical measure of how much the sentence: 1) conveys the image content (relevance) in terms of the objects, actions and scene predicted and 2) is grammatically correct (readability). We engaged the services of Amazon Mechanical Turks (AMT) to judge the generated sentences based on a discrete scale ranging from 1–5 (low relevance/readability to high relevance/readability). The averaged results of ROUGE-1, R_1 and mean length of the sentences with the Relevance+Readability scores for all experiments are summarized in Table 3. For comparison, we also asked the AMTs to judge the ground truth sentences as well.

4.2 Discussion

The results reported in Table 3 reveals both the strengths and some shortcomings of the approach which we will briefly discuss here. Firstly, the R_1 scores indicate that based on a purely summarization (unigram-overlap) point of view, the proposed approach of using the HMM to predict $\mathcal{T}^*$ achieves the best results compared to all other approaches with $R_1 = 0.44$. This means that our sentences are the closest in agreement with the human annotated ground truth, correctly predicting the sentence structure components. In addition sentences generated by $\mathcal{T}^*$ are also succinct: with an average length of 6.9 words per sentence. However, we are still some way off the human gold standard since we do not predict other parts-of-speech such as adjectives and adverbs. Given this fact, our proposed approach performance is comparable to other state of the art summarization work in the literature [Bonnie and Dorr, 2004].

Next, we consider the Relevance+Readability metrics based on human judges. Interestingly, the first baseline, $\mathcal{T}^*_{b1}$ is considered the most relevant description of the image and the least readable at the same time. This is most likely due to the fact that this recall oriented strategy will almost certainly describe some objects but the lack of any verb description; and longer sentences that average 8.2 words per sentence, makes it less readable. It is also possible that humans tend to penalize less irrelevant objects compared to missing objects, and further evaluations are necessary to confirm this. Since $\mathcal{T}^*_{b2}$ is limited to two objects just like the proposed HMM, it is a more suitable baseline for comparison. Clearly, the results show that adding the HMM to predict the optimal sentence structure increases the relevance of the produced sentence. Finally, in terms of *readability*, $\mathcal{T}^*$ generates the most readable sentences, and this is achieved by leveraging on the corpus to guide our predictions of the most *reasonable* nouns, verbs, scenes and prepositions that agree with the detections in the image.

5 Future Work

In this work, we have introduced a computationally feasible framework that integrates visual perception together with semantic grounding obtained from a large textual corpus for the purpose of generating a descriptive sentence of an image. Experimental results show that our approach produces sentences that are both relevant and readable. There are, however, instances where our strategy fails to predict the appropriate verbs or nouns (see Fig. 9). This is due to the fact that object/scene detections can be wrong and noise from the corpus itself remains a problem. Compared to human gold standards, therefore, much work still remains in terms of detecting these objects and scenes with high precision. Currently, at most two object classes are used to generate simple sentences which was shown in the results to have penalized the relevance score of our approach. This can be addressed by designing more complex HMMs to handle larger numbers of object and verb classes. Another interesting direction of future work would be to detect *salient* objects, learned from training image+corpus or eye-movement data, and to verify if these objects aid in improving the descriptive sentences we generate. Another potential application of representing images using $\mathcal{T}^*$ is that we can easily sort and retrieve images that are similar in terms of their *semantic* content. This would enable us to retrieve, for example, more relevant images given a verbal search query such as $\{$`ride,sit,fly`$\}$, returning images where these verbs are found in $\mathcal{T}^*$. Some results of retrieved images based on their verbal components are shown in Fig. 10: many images with dissimilar visual content are correctly classified based on their semantic meaning.

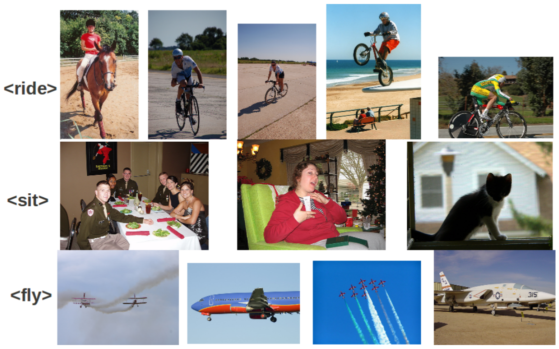


Figure 10: Images retrieved from 3 verbal search terms: `ride,sit,fly`.

6 Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1035542. In addition, the support of the European Union under the Cognitive Systems program (project POETICON) and the National Science Foundation under the Cyberphysical Systems Program, is gratefully acknowledged.

References

Berg, T. L., Berg, A. C., Edwards, J., and Forsyth, D. A. (2004). Who's in the picture? In *NIPS*.

Bonnie, D. Z. and Dorr, B. (2004). Bbn/umd at duc-2004: Topiary. In *In Proceedings of the 2004 Document Understanding Conference (DUC 2004) at NLT/NAACL 2004*, pages 112–119.

Choi, J. D. and Palmer, M. (2010). Robust constituent-to-dependency conversion for english. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories*, pages 55–66, Tartu, Estonia.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2008). The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results.

Farhadi, A., Hejrati, S. M. M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. A. (2010). Every picture tells a story: Generating sentences from images. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pages 15–29. Springer.

Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. (2008). Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645.

Golland, D., Liang, P., and Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proceedings of EMNLP*.

Graff, D. (2003). English gigaword. In *Linguistic Data Consortium, Philadelphia, PA*.

Jie, L., Caputo, B., and Ferrari, V. (2009). Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. In NIPS, editor, *Advances in Neural Information Processing Systems*, NIPS. NIPS.

Kojima, A., Izumi, M., Tamura, T., and Fukunaga, K. (2000). Generating natural language description of human behavior from video images. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 4, pages 728 –731 vol.4.

Kourtzi, Z. (2004). But still, it moves. *Trends in Cognitive Sciences*, 8(2):47 – 49.

Liang, P., Jordan, M. I., and Klein, D. (2009). Learning from measurements in exponential families. In *International Conference on Machine Learning (ICML)*.

Lin, C. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACLHLT*.

Lin, H.-T., Lin, C.-J., and Weng, R. C. (2007). A note on platt's probabilistic outputs for support vector machines. *Mach. Learn.*, 68:267–276.

Mann, G. S. and Mccallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. In *The 24th International Conference on Machine Learning*.

McKeown, K. (2009). Query-focused summarization using text-to-text generation: When information comes from multilingual sources. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, page 3, Suntec, Singapore. Association for Computational Linguistics.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Schwartz, W., Kembhavi, A., Harwood, D., and Davis, L. (2009). Human detection using partial least squares analysis. In *International Conference on Computer Vision*.

Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A. (2003). Context-based vision system for place and object recognition. In *ICCV*, pages 273–280. IEEE Computer Society.

Traum, D., Fleischman, M., and Hovy, E. (2003). Nl generation for virtual humans in a complex social environment. In *In Proceedings of he AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, pages 151–158.

Urgesi, C., Moro, V., Candidi, M., and Aglioti, S. M. (2006). Mapping implied body actions in the human motor system. *J Neurosci*, 26(30):7942–9.

Yang, W., Wang, Y., and Mori, G. (2010). Recognizing human actions from still images with latent poses. In *CVPR*.

Yao, B. and Fei-Fei, L. (2010). Grouplet: a structured image representation for recognizing human and object interactions. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA.

Yao, B., Yang, X., Lin, L., Lee, M. W., and Zhu, S.-C. (2010). I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485 –1508.