Understanding Images with Natural Sentences

Yosihitaka Ushiku
Grad. School of Information Science and Technology
The University of Tokyo
ushiku@isi.imi.i.u-tokyo.ac.jp

Tatsuya Harada
Grad. School of Information Science and Technology
The University of Tokyo / JST PRESTO
harada@isi.imi.i.u-tokyo.ac.jp

Yasuo Kuniyoshi
Grad. School of Information Science and Technology
The University of Tokyo
kuniyosh@isi.imi.i.u-tokyo.ac.jp

ABSTRACT

We propose a novel system which generates sentential captions for general images. For people to use numerous images effectively on the web, technologies must be able to explain image contents and must be capable of searching for data that users need. Moreover, images must be described with natural sentences based not only on the names of objects contained in an image but also on their mutual relations. The proposed system uses general images and captions available on the web as training data to generate captions for new images. Furthermore, because the learning cost is independent from the amount of data, the system has scalability, which makes it useful with large-scale data.

Categories and Subject Descriptors

I.2.10 [Artificial Intelligence]: Vision and Scene Understanding

General Terms

Algorithms, Theory

Keywords

Similarity measure, Probabilistic canonical correlation analysis, Multi-stack decoding

1. WHY SENTENCES?

A system by which a computer understands contents of multimedia automatically is indispensable for use with many multimedia web contents effectively. Existing methods of describing multimedia contents express contents with labels of objects, which is insufficient. Description with natural sentences including the objects’ relations is important. For example, existing image search engines, which search for required images based on an example image or some words, cannot understand the objects’ relations. Therefore, the engine cannot distinguish a sentence such as “A person bit a dog.” or a picture in which a person bit a dog from data in which a dog bit a person because the engine can only understand independent words (Fig.1).

In other methods, the objects contained in each image are described using some labels to support searches and a user’s understanding of images and videos in general web sites and sharing services. Although visual recognition, which has been studied actively, can perform this automatically, users cannot understand a story including relations between objects based on labels alone. Automatic generation of a multimedia caption enables users to search for data with a more flexible query and to understand the contents of each datum more easily.

2. DIFFICULTIES AND SOLUTION

Automated captioning of images is extremely difficult because it necessitates correct estimation not only of the objects in the images, but also correct estimation of those objects’ mutual relations and expressions using correct grammar. For example, one-to-one relations of a thousand objects can run into the millions. Prior research of sentence outputs from a video, character recognition of subtitles, and speech recognition of conversation can be used instead of descriptions of the visual contents.

Moreover, the problem exists of what kinds of data are used for training to generate captions. Large quantities of training data are necessary to generate captions of general images. Therefore, using a vast quantity of web data, as is done for existing image recognition, presents the most realistic approach. However, only a single study [2] of caption generation uses triplets expressing (object, action, scene) for an image and a caption. It costs users dearly because it necessitates manual labeling.

In [8], we proposed caption generation for general images using data consisting only of images and texts. We devote
Attention to images, the contents of which resemble the contents of an inputted image, and persuade as follows.

Assumption 1. In captions of similar images, some representations explain some parts of an inputted image. The objects in an image, their mutual relations and grammatical knowledge are already embedded there. Consequently, the image contents are explainable both grammatically and correctly by reconstructing these captions.

Therefore, captions are generated by (i) searching for similar training samples, and by then (ii) reconstructing the retrieved captions for an input image (Fig.2).

In searching for similar training samples, images with captions that partly explain the inputted image should be retrieved. Briefly, the objective is the acquisition of training samples with contents resembling the input image. Generally, however, a gap separates the similarity of image appearance and the similarity of contents. Therefore, we proposed the improvement of the gap by combining the similarity of an image and the similarity of a text [7]. In that study, a dataset consisting of pairs of an image and a text is constructed from web contents. Then the precision of searching for images with contents resembling the input image is superior to that obtained when using the similarity of visual appearance. Moreover, the approach is good for large-scale data because the computational cost for similarity learning is independent of the amount of data.

Judging which portion of each sentence explains the input is necessary for generation of a caption of an input image by reconstructing the captions of the searched images. Therefore, each sentence is divided into phrases of several words. Then phrases with following two properties are scored highly as discriminative, i.e. explainable phrases.

- It often appears in the captions of searched images.
- It seldom appears in the captions of other images.

A caption combining those phrases is generated to maximize the sum of the scores.

3. DEMO

This demonstration is based on the technique proposed in [8]. Two functions are executable: (a) caption generation of an input image (Fig.3) and (b) image search from an input sentence. Users can search for images using a text query. The automatically generated caption is displayed on each image of search results. In this demonstration, learning is performed with the dataset consisting of images and texts offered in [2]. A caption is generated for new images.

First, (a) caption generation for an input image is described. The application has finished learning for similar image searches, according to results of our earlier study [8]. When a new image is inputted, similar images are sought, as in Sec. 2. The captions of the searched images are decomposed into arbitrary phrases, and a caption is reconstructed.

In [8], we do not perform (b) image search from an input sentence directly. Although it is necessary to assign a caption to an image beforehand for searching, variation occurs in vocabulary and word order when expressing an image in a natural sentence. In other words, even if the query sentence differs from the captions, some images have the same contents as those of the query. Then, the score is saved with a phrase that was not used at the time of caption generation, and the variation in expressions is absorbed by matching with the arbitrary phrases of an input query. The larger the scores of matched phrases become, the more highly the image is ranked among search results.

Because the dataset is small, the explainable contents are restricted. However, even in a larger dataset, caption accuracy is improved because similar images will accompany various sentences. The proposed technique can accept larger dataset because it is scalable, as described in Sec. 2.

4. CONCLUSION

As described in this paper, automatic caption generation for images is proposed as a new form of image understanding. The expressions in a sentence show not only the names of the objects; they also express their respective relations, which eases users' comprehension. Therefore, this method is effective for use with large-scale multimedia. Especially, this method is useful for seeking desired data more efficiently from many images, and for explaining the contents of each image automatically. Because of these benefits, the method is particularly useful for general websites and personal devices such as cameras and mobile phones. Furthermore, for data which require a long time for viewing such as videos and image folders, such caption generation is useful because users
can search for desired multimedia with natural sentences and experience the contents without playbacking. Therefore, caption generation for use with multimedia—not only for single images but also for groups of images and a video—is the subject of our future work.

5. REFERENCES


APPENDIX

In this section, we present the details of the proposed approach and evaluations. We describe them to make this paper self-contained though they are also described in [8].

A. SENTENTIAL CAPTION GENERATION

Sec. A.1 describes similarity learning used to search for similar training samples, and Sec. A.2 describes caption generation for an input image from the retrieved captions.

A.1 Retrieving training samples

Generally, a gap separates the similarity of image appearance and the similarity of contents. To search for training samples with contents resembling the input image from large data, scalable similarity learning must be done. Ushiku et al. [7] proposed the improvement of the gap by combining the similarity of an image and the similarity of a text. This method is an extension of the image annotation method Canonical Contextual Distance (CCD) [3] to images and texts. The algorithm can calculate the similarity between images as a distance of that context. Then we can search more precisely than when using the similarity of visual appearance. Moreover, the method is scalable because this similarity learning is independent of a data number.

We denote an image/text feature as \textbf{z} or \textbf{y}, and the latent variable as \textbf{z}. It is realized that \textbf{z} or \textbf{y} has appeared as a stochastic representation of an image or a text from \textbf{z} based on a certain contents. Then, a similar image retrieval turns into a nearest neighbor approach using the distance between probability distributions \( p(\mathbf{z|x,y}) \) from an input image and \( p(\mathbf{z|x,y}) \) from the training sample pair of an image and a text. These distributions are expressed as a normal distribution as \( \mathbf{z|x,y} \sim N(\mu_{xy},\Phi_{xy}) \). For an exponential family such as a normal distribution, Kullback–Leibler Divergence (KLD) is used as a pseudo-distance. We denote i.i.d. \( p(\mathbf{z}|\mathbf{x}) \) with mean \( \mu_{x} \) and with variance \( \Phi_{x} \), and \( p(\mathbf{z}|\mathbf{x,y}) \) with mean \( \mu_{xy} \) and variance \( \Phi_{xy} \). Then CCD is defined by excluding constant terms from KLD from an input image \( \mathbf{x}_{q} \) to a training sample \( \{\mathbf{x},\mathbf{y}\} \) as

\[
\text{CCD}(\mathbf{x}_{t},\mathbf{y}_{t}|\mathbf{x}_{q}) = \left\| \Phi_{xy}^{-\frac{1}{2}} (\mu_{xy} - \mu_{x}) \right\|^2. \tag{1}
\]

When we define latent variables \( r_{t} \) for a training sample and \( q_{t} \) for the input image, this is equivalent to a neighbor search using Euclidean distance with \( r \) defined as

\[
r_{t} = \Phi_{xy}^{-\frac{1}{2}} \mu_{xy}, q_{t} = \Phi_{xy}^{-\frac{1}{2}} \mu_{x}. \tag{2}
\]

As described above, it is reduced to the estimation of \( \mu_{xy} \) and \( \Phi_{xy} \). If maximum likelihood estimation is performed with training samples, then it is derived as

\[
\mu_{xy} = M_{x}^{-1}A^{T}(x - \bar{x}), \Phi_{xy} = I - M_{x}M_{y}^{T} . \tag{3}
\]

\[
\mu_{xy} = \left( \frac{M_{x}^{-1} A^{T} (I - \Lambda)^{-1} (I - \Lambda)^{-1} \Lambda (I - \Lambda)^{-1} A}{\Phi_{xy}} \right) , \tag{4}
\]

\[
\mu_{xy} = I - \frac{M_{x}^{-1} A^{T} (I - \Lambda)^{-1} (I - \Lambda)^{-1} \Lambda (I - \Lambda)^{-1} A}{\Phi_{xy}} , \tag{5}
\]

where \( \bar{x} \) and \( \bar{y} \) are feature averages, and where \( M_{x} \) and \( M_{y} \) are arbitrary matrices satisfying \( M_{x}M_{y}^{T} = \Lambda \) and for which the spectrum norm of each matrix is less than 1. Here we denote them with \( 0 < \beta < 1 \) as \( M_{x} = \Lambda^{\beta} \) and \( M_{y} = \Lambda^{1-\beta} \), which is exactly like probabilistic canonical correlation analysis (pCCA): a probabilistic interpretation of canonical correlation analysis (CCA). In fact, CCA maximizes the correlation between canonical variables \( s = A^{T}(x - \bar{x}) \) and \( t = B^{T}(y - \bar{y}) \). These are derived from the following generalized eigenvalue problem: \( R_{XY}R_{XY}^{T} = R_{XY}AA^{T}, R_{XY}R_{XY}^{T} = R_{XY}BB^{T} \), and \( R_{XY} = \sum_{i=1}^{n}(x_{i} - \bar{x})(x_{i} - \bar{x})^{T}, R_{Y} = \sum_{i=1}^{n}(y_{i} - \bar{y})(y_{i} - \bar{y})^{T}, R_{XY} = \sum_{i=1}^{n}(x_{i} - \bar{x})(y_{i} - \bar{y})^{T} = R_{Y}^{1/2}X^{1/2} \).

A.2 Reconstructing training sample captions

We define the training sample pair \( s_{t} \) of an image \( x_{t} \) and a caption \( y_{t} \), caption set of \( s_{t} \) as \( W_{t} \), and each caption as \( w_{t} = w_{t1}w_{t2} \ldots w_{ti} \), where \( l \) is the length and \( w_{ti} \) represents the \( i \)-th word. Here, \( s_{t} \) is obtained using a neighborhood search with CCD described in the preceding section. Some phrases (word sequences) that explain the input image’s own objects and those relations correctly and grammatically are expected in \( W_{t} \).

We generate captions of desired length \( l_{tar} \) to the input image. The scores of candidate captions are computed from the input image; then the caption having the highest score is output. Our objectives are 1) searching for explanatory phrases and 2) captioning for the desired length. Consequently, two feature functions are defined for each purpose. The first is a score \( \phi(p_{hr}(s_{t}|x_{q})) \) when phrase \( p_{hr} \) is selected for the input image \( x_{q} \). The second is a score for the current sentence length. Although other scores such as those of a grammar model with n-gram are usable, here we investigate the captioning ability accomplished solely by reconstruction of training sample captions.

We define the probability of retrieving \( s_{t} \) from an input image \( x_{q} \) as \( p(s_{t}|x_{q}) \), and define the probability of selecting a phrase \( p_{hr} \) from the pair \( x_{q} \) as \( p(p_{hr}|s_{t}) \). The phrase score is defined as a logarithm with those probabilities as

\[
\phi(p_{hr}|s_{t}|x_{q}) = \left( 1 - \frac{\alpha(p_{hr})}{\alpha_{max}(p_{hr})} \right) \log \left( \sum_{s_{t}} p(p_{hr}|s_{t})p(s_{t}|x_{q}) \right) , \tag{6}
\]

where \( \alpha_{max}(p_{hr}) \) is the maximum occurrence of \( p_{hr} \) in one sentence, and \( \alpha(p_{hr}) \) is the occurrence of \( p_{hr} \) in the currently generating sentence. The coefficient of the logarithm has a control for a specific dominant phrase so as not to produce meaningless repetition. The probability \( p(s_{t}|x_{q}) \) of selecting a pair \( s_{t} \) is retrieved from the input image \( x_{q} \) is defined as \( p(s_{t}|x_{q}) = \exp(-CCD(s_{t}|x_{q}))/\sum_{s_{t}} \exp(-CCD(s_{t}|x_{q})) \). The probability \( p(p_{hr}|s_{t}) \) of selecting \( p_{hr} \) from each pair \( s_{t} \) is defined using tf-idf [6], the product of the frequency of each term in a document and...
the inverse of the appearance frequency of that term in all documents, as \( p(p_h|s_t) = \text{tf-idf}(p_{hr}) / \sum_{p_{hr} \in V} \text{tf-idf}(p_{hr}) \). Therein, TF-idf represents how well each word distinguishes the text.

We denote the length score as a logarithm of the probability of current length \( l \sim N(l_{tar}, \sigma) \). Qualitatively, \( \sigma \) represents the tolerance level of caption length. In this paper, we define that \( l_{tar} = 12 \) and \( \sigma = 1 \).

Because generable captions increase exponentially with vocabularies, comparison of all captions is difficult. Consequently, output is determined with graph search. In a typical approach [4] to statistical machine translation, scores are first calculated using various feature functions from both languages. Then the translations are generated to maximize the sum of the scores using a graph search such as a multi-stack beam search.

Beam searching saves some high rank candidate sentences in one stack. The stack size can adjust the accuracy of graph search and the computational complexity. In a multi-stack beam search, a stack is divided according to the translated word amounts in the input sentence. The highest-ranked sentence in the last stack is output. In this paper, however, the input not as a sentence but as an image causes problems.

First, because there is nothing equivalent to word position correspondence of an input sentence and a translation, left-to-right expansion is impossible. Therefore, we take an approach to expand a sentence to both sides. We consider the \( l_s \) words’ phrase which has the highest probability of explaining the input image as a seed. Then we expand a caption to both sides.

As another problem, because the judgment of untranslated portion of an image is difficult, neither the end of the translation and division of the stack based on the translated words’ length is possible. Therefore, we divide a stack based on the sentence being generated. An index on which can be ended is given to the candidate captions in each stack. In the caption set \( W \), EDGE of Sentence (EOS) is added beforehand to the head and end of each caption. When the phrase that is connectable to EOS exists in \( W \), it is unrelated with the score when connected with EOS at the time, but the index and the score when connected to the EOS are memorized. When the expansion is finished, the word sequence that is inserted by EOS will be output as

\[
\omega = \arg \max_{w=(w_{i_s}, w_{i_s}+1, \ldots, w_{i_e}-1, w_{i_e})} \sum_k \phi_k (w_{i_s}^+ | x_q), \tag{7}
\]

where \( i \) is an index connectable to EOS, \( i_e \) and \( i_s \) are found in left-to-right and right-to-left expansion respectively.

**B. EVALUATION**

In [2], image captions themselves are manually evaluated, which makes fair comparison difficult. Therefore, we investigate automatic evaluation with which matching to the reference given to the input image in the dataset is quantified. Captions are evaluated with match of n-gram using BiLingual Evaluation Understudy (BLEU) [5] and NIST [1].

The CCD parameter used to search for a training sample is optimized by maximizing BLEU/NIST of the caption which originally accompanied the input image and the retrieved caption. Concurrently, Mean Average Precision (MAP) is also used as a typical retrieval evaluation. This calculates the match rate of the category of an input image and the retrieved sample. However, this evaluation can not be calculated without the category label to images and captions.

For caption generation from images and captions only, optimization without MAP, which needs labels, is desirable. Automatic evaluation was compared with evaluation by hands in advance of evaluating the proposed method. Six captions output to each image are shown to 10 subjects. The order of captions is random. For that reason, subjects cannot understand which system output which caption. Each caption is evaluated using a five-level rating according to how well it explains the image. The evaluations are standardized. The correlations between each of BLEU and NIST and hand evaluation are 0.63 and 0.82, respectively, showing that NIST especially is appropriate for evaluation of captions.