Generating Sentences From Abstract Images: ECE 5554 Project Report

Shanmukha Ramakrishna Vedantam
Virginia Tech
vrama91@vt.edu

Abstract

Both language and vision are fundamental to perception in human beings. Language has been shown to boost invisible objects into visual awareness. Understanding this connection can help us solve vision problems better. A first approach for this is to generate meaningful and coherent natural language descriptions of images. In this project, various methods for describing images, both using text corpora associated with images and using only native image content are explored and compared on a metric that measures human agreement. Five methods are compared and evaluated on two metrics and the results discussed.

1. Introduction

Generating semantically meaningful descriptions of images is a major step towards semantic scene understanding. Enabling machines to generate such descriptions opens up multiple avenues for richer human-machine interaction and communication. Relating natural language to image content is also very useful for mining information about in web. Recent work has shown that language can even boost objects into visual perception [3].

Other work [5] has shown that photorealism is not necessary for the study of semantic understanding. This work introduces a novel abstract scene “clip art” dataset. This dataset is used in the project for generating the sentences. Using clip art instead of normal images gives us the advantage of not having to rely on noisy image attribute and object detectors and focus better on capturing higher level semantic aspects.

Approaches for image to text have used caption transfer from a large collection of Images [4] and combining object detections and attribute detections with language statistics or spatial relationships for preposition prediction [2].

In this project, various methods for image to text, including those using text features and visual features are explored and compared against one another. Using text features gives us a rough upper bound on how well a good visual features based approach can perform on our evaluation metric, since the text corpora used to model the descriptions are human generated. Specifically, the methods used are direct bag of words transfer to generate sentences and first and fourth order markov chains modelled on thirty human descriptions of per image. Visual features such as spatial envelopes (GIST) and some semantic occurence features described section 2 are used for sentence transfer using a one nearest neighbour approach.

2. Approach

The five methods used to generate sentences from clip-art images are described in this section. The bag of words sentence generation, nearest neighbour based sentence transfer, computation of semantic occurence features and template based sentnece generation codes have been implemented by me. Implementation for markov chain based sentence generation was taken from - https://github.com/hrs/markov-sentence-generator. A wrapper script was written over this implementation to generate sentences for the clip-art dataset.

2.1. Bag of Words

A bag of words representation is taken for each clip art image from the dataset of [5]. These bag of words are generated by first getting user descriptions for the 1002 images in the dataset. From this corpus, individual words are extracted to form a dictionary representation, of which the top 355 words are retained. This is the dictionary space. Given a test image, the 355 length dictionary representation for the image is read and the corresponding words are copied off into a sentence, unordered. This forms our Bag of Words sentence generation baseline.

2.2. Markov Chains

The second type of text based generation scheme used is to model Markov Chains. In an n-gram markov model, the next word is predicted based on the probability of occurence of some word conditioned on the four previous words being the ones currently present in the sentence. These probabilities are computed from a corpus of human descriptions collected a subset of clipart images (test images).

Two types of markov chains are used for sentence generation: first order and fourth order. The first order markov chain preserves the pairwise dependencies between words from the statistics of a corpus. The fourth order markov chain estimates the next word based on the current word and the three previous words. The seed n - gram is chosen at random from the training corpus. To ensure that sentences end, start sentence[s] and end sentence[/s] tokens are embedded into the corpus.

There are two key design choices for generating sentences using markov chains. First is the choice of the model corpus. Firstly, I considered using all the 30 descriptions for each of the 60 images to build a corpus of 1800 sentences and to train a corpus using those sentences. However, that gives us no way to incorporate identity of the image into the sentence generation process. Thus each image is trained only with its ground truth human descriptions. The idea behind doing markov chains is to jumble up the ordering of words within and across sentences to create poor copies of the original ground truth sentence. The second design choice is the value for the parameter "n" where n is the order of the markov chain. 1 and 4 are chosen since the first order chain represents the lowest level of dependencies and should hence be overall of a much poorer quality than the fourth order markov chain in terms of semantics. Thus it is expected that there will be significant qualitative and quantitative differences between the two.

2.3. Nearest Neighbour using GIST

The GIST descriptor, of length 512 is extracted for each of the test images. The test descriptor is then matched to the descriptors of the train images. Train images are those in the dataset for which ground truth descriptions are already known. We use image features to find the closest training image whose description is known and transfer the description to our test image.

2.4. Nearest Neighbour using Semantic Occurence Features

A semantic occurence feature space is defined for nearest neighbour using object occurence statistics for each test and train clipart image. The object occurence statistics list the names of various atomic images such as boy in various poses and expressions, bear, crown etc. used in the final clip art image. These detail all the instances of every category. Apart from binary occurence, the statistics also inform us about the location (x, y, relative depth), orientation (left/right), category (sky, large object, boy, girl, animal, clothes, food, toys).

The key choice here is the selection of the feature space for nearest neighbours. A hierarchical feature space is chosen. We match both instance level occurences and a subset of category level occurences. The "boy" and "girl" category are chosen as features since both these categories are semantically dense. That is, there are a lot of very similar but slightly different images in these two categories. Thus a candidate nearest neighbour match should not be penalized too much for failing to distinguish exactly between an angry mike and a confused mike, for instance. This is as against failing to distinguish between a bear and a dog, both of which belong to the animal category. The semantic features have 126 instance level occurence features which are concatenated with two category level occurence features taking the feature vector to 128 dimensions.

Nearest Neighbour search is performed in this feature space as described in the previous subsection and sentences are transferred to the test instance.

2.5. Template Based Sentence Generation

I now describe the template based sentence generation which uses the semantic occurence statistics to generate novel sentences using image statistics. This is done by first extracting all the instance level objects in a scene from the occurence statistics and endowing them with manually chosen attributes. For instance, an occurence of Mike(Boy) in a squatting position with a smile on his face is mapped to the word sequence "happy, squatting mike". Some objects categories such as "toys" are endowed with the verb "playing". Thus the occurence of a soccer ball is mapped to the word sequence "playing soccer". This is done for the categories "clothes" where the verb is wearing and "food" where the verb is "eating".

Once this mapping is done for each instance level occurence, we now need to figure out who the doers of the actions are. We use the location information of the "boy" instance , "girl instance" and "clothes", "food" and "toys" to figure out for example, if the boy is "eating" or the girl or both. If the difference in the distance between boy and the object on one hand and girl and the object on the other is below a threshold (set to 10) we declare that both are doing the activity, else we either say mike is doing the activity or jenny. Finally, we fit all the detections into the following template:

1. mike_sent= [adjectives,mike] is [playing,toy-instance] and [eating,food-instance] and [wearing, clothes-instance]

2. jenny_sent = and [adjectives, jenny] is [playing, toy-instance] and [eating, food-instance] and [wearing, clothes-instance]

3. both_sent = and they are [playing,toy-instance] and [eating, food-instance] and [wearing, clothes-instance]

4. last_sent = [in front of/alongside] [skyobject / large object] [while,animal-instance,watches on].

5. Each item enclosed in tags [] is optional, that is if an association is found between mike and some instance level occurence only then we include it in the sentence for mike.

6. The final output sentence of the system is a concatenation of mike_sent , jenny_sent, both_sent and last_sent, with appropriate spacing inserted.

3. Experiments and results

The experimental setup and the evaluation metrics, as well as qualitative and quantitative results are detailed below.

3.1. Dataset

The Dataset used is the clipart dataset introduced by [5] which contains 1002 clipart images. For each of the image, this dataset provides word features, namely bag of words on a 355 sized dictionary, visual features which detail the relative spatial occurences and mutual information of object category / instance with class/semantic uniqueness. It also provides object occurence statistics, which can be used to exactly reconstruct any clip-art image. The object occurence statistics include instance level names of the object, location in x, y and spatial depth, category (0 to 7) and flip (orientation: left or right). The test images for which sentences are generated using my approaches are a randomly chosen subset of 60 images. The dataset is augmented by collecting 30 human sentences for each of these images. These sentences are then fit into the evaluation metric framework to judge the qualitiy of a generated sentence. The evaluation metric is described in detail in the next section.

3.2. Evaluation Metric

The key idea behind my evaluation metric is to capture the consensus or a human agreement behind what an image is supposed to express. Individual human generated sentences might not be too reliable for matching, especially because people are very creative at such tasks as describing clipart images, but if we can successfully capture the relevant averaging behaviour, we can call candidate sentence a good description of the image if there is an acceptable overlap between the average behaviour of the corpus ground truth human sentences and the test generated sentence. A formalization of this intution is TF-IDF (Term Frequency and Inverted Document Frequency weighting) on which we apply LSI(Latent Semantic Indexing) [1] to produce the evlauation metric. The evaluation is not only done at a one gram level but also using a two gram dictionary.

The framework for evaluation of a test sentence is now described. We find the TFIDF - LSA similarity of the test sentence to each of the test images by matching with the human descriptions for images. If we retrieve the ground truth image, for which the sentence generation task was performed, we call that a match. We refer to the number of matches as the accuracy of the metric.

3.3. Qualitative Results

We now take an image from our test dataset and see the performance of the various methods on the test image.

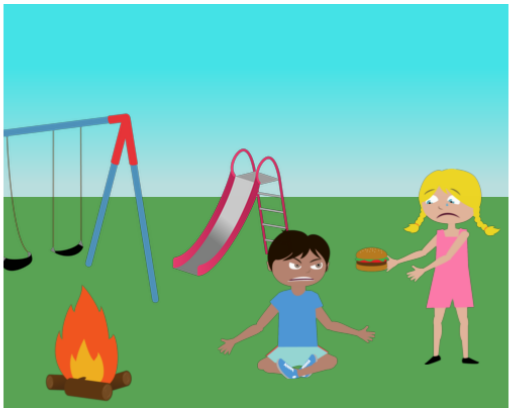

Figure 1. Sample Image from the test dataset

- Bag Of Words:mike jenny and to her that she play pie pizza because has not eat can friend ready shout
- 1 - Order Markov Chain: Jenny.
- 4 - Order Markov Chain: Mike is serving Jenny a pie and pizzajenny is yelling while Mike holds up a pie and a pizza.
- GIST + NN:Mike and Jenny are both angry while the cat is sitting next to Jenny who is wearing a witches hat and Mike who is holding a tennis racket while there is a lightening bolt in the background.
- Semantic Features + NN: Jenny's pizza smelled so good that Mike couldn't wait to taste it.
- Template Generated: smiling mike is eating pie and eating pizza and wearing chef's hat and wearing sunglasses and surprised jenny alongside a tree.

3.4. Results

We now see how the method performs with respect to two evaluation criteria: 1-gram TFIDF + LSA and 2-gram TFIDF + LSA. The n in n-gram essentially signifies the number of words contained in the dictionary representation. As n-grams increase, the size of the dictionary increases as the number of representable combinations grows. Also, as n increaeses there is greater sparsity in the representation

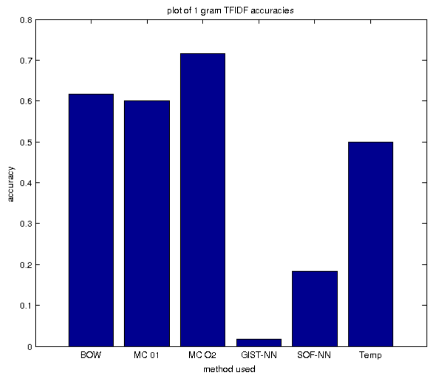

Figure 2. Result for 1-gram TFIDF. X axis shows the method that is used 1) BOW 2) Markov Chain Order 1 3) Markov Chain Order 4 4)GIST + Nearest Neighbour 5) Semantic Occurence Features + Nearest Neighbour 6) Template Sentence Generation

and chances are that with more sentences, the representation gets better. Presently both the "n - grams" are trained on the same number of sentences (30).

Clearly, we see that the word based methods, have an upper hand since they have access to ground truth human generated words. They only make mistakes in the ordering and miss out on selecting somewords (markov chains), but generally do fairly well. The dictionary basis for the evaluation set of 30 images and the BOW test sentences is different. This explains why BOW doesn't perform as well as one would expect. Further we note that Semantic Features + Nearest Neighbour performs better than GIST + Nearest Neighbour. This shows that GIST is not a good feature to infer semantics on our dataset. Template based sentence generation seems to outperform all other image feature based sentence generation methods.

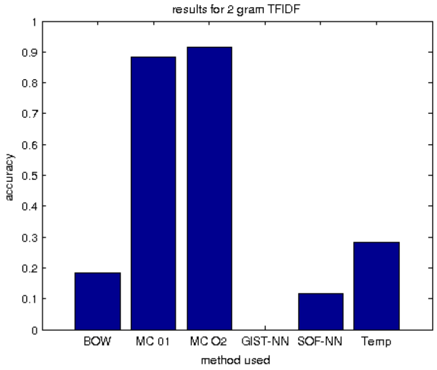

Figure 3. Result for 2-gram TFIDF. X axis shows the method that is used 1) BOW 2) Markov Chain Order 1 3) Markov Chain Order 4 4)GIST + Nearest Neighbour 5) Semantic Occurence Features + Nearest Neighbour 6) Template Sentence Generation

The results for the 2 gram TFIDF are more or less consistent with what we saw for 1 gram. Template generated sentences are still the best performing method when it comes to using image features. The score for both 1 st order and 4th order markov chains are pretty high as expected, since both 1st order and 4th order markov chains preserve pairwise co-occurences.

As a baseline, the detections for template based generation were made random and sentences were then constructed. The results for 1 - gram accuracy using this baseline comes out to be 0.0167 on a scale of 1.

4. Conclusion and Future Work

Generating natural language from image content is a difficult problem. This project attempted to try out a few baseline approaches for image to text on clip art images. The template based sentence generation scheme looks promising. The connecting words between categories/objects are presently hard coded. For instance [and] is inserted between any two objects like food or clothing. These words can be inferred from a language model instead. Further the dataset of [5] has a number of features such as co-occurence, mutual information between classes, distribution of positions of objects and its relation to prepositional usage and so on that have not been used for this project. If these are leveraged, the sentence generation can be made to look more natural.

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

[2] T. B. Kulkarni, Sagnik Dhar. Babytalk: Understanding and generating simple image descriptions, 2011. In Proc. CVPR.

[3] G. Lupyana and E. J. Wardb. Language can boost otherwise unseen objects into visual awareness,pnas, 2013. http://www.pnas.org/content/early/2013/08/08/1303312110.

[4] T. L. B. Vicente Ordonez, Girish Kulkarni. Im2text: Describing images using 1 million captioned photographs, 2011. In Proc. NIPS.

[5] C. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction, 2013. In CVPR.