Using Maximum Entropy for Automatic Image Annotation

Jiwoon Jeon and R. Manmatha

Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts Amherst
Amherst, MA-01003.
{jeon, manmatha}@cs.umass.edu,

Abstract. In this paper, we propose the use of the Maximum Entropy approach for the task of automatic image annotation. Given labeled training data, Maximum Entropy is a statistical technique which allows one to predict the probability of a label given test data. The techniques allow for relationships between features to be effectively captured. and has been successfully applied to a number of language tasks including machine translation. In our case, we view the image annotation task as one where a training data set of images labeled with keywords is provided and we need to automatically label the test images with keywords. To do this, we first represent the image using a language of visterms and then predict the probability of seeing an English word given the set of visterms forming the image. Maximum Entropy allows us to compute the probability and in addition allows for the relationships between visterms to be incorporated. The experimental results show that Maximum Entropy outperforms one of the classical translation models that has been applied to this task and the Cross Media Relevance Model. Since the Maximum Entropy model allows for the use of a large number of predicates to possibly increase performance even further, Maximum Entropy model is a promising model for the task of automatic image annotation.

1 Introduction

The importance of automatic image annotation has been increasing with the growth of the worldwide web. Finding relevant digital images from the web and other large size databases is not a trivial task because many of these images do not have annotations. Systems using non-textual queries like color and texture have been proposed but many users find it hard to represent their information needs using abstract image features. Many users prefer textual queries and automatic annotation is a way of solving this problem.

Recently, a number of researchers [2, 4, 7, 9, 12] have applied various statistical techniques to relate words and images. Duygulu *et al.* [7] proposed that the image annotation task can be thought of as similar to the machine translation problem and applied one of the classical IBM translation models [5] to this problem.

Jeon *et al.* [9] showed that relevance models (first proposed for information retrieval and cross-lingual information retrieval [10]) could be used for image annotation and they reported much better results than [7]. Berger *et al.* [3] showed how Maximum Entropy could be used for the machine translation tasks and demonstrated that it outperformed the classical (IBM) machine translation models for the English-French translation task. The Maximum Entropy approach has also been applied successfully to a number of other language tasks [3, 14].

Here, we apply Maximum Entropy to the same dataset used in [7, 9] and show that it outperforms both those models. We first compute an image dictionary of visterms which is obtained by first partitioning each image into rectangular regions and then clustering image regions across the training set. Given a training set of images and keywords, we then define unigram predicates which pair image regions and labels. We automatically learn using the training set how to weight the different terms so that we can predict the probability of a label (word) given a region from a test image. To allow for relationships between regions we define bigram predicates. In principle this could be extended to arbitrary n-grams but for computational reasons we restrict ourselves to unigram and bigram predicates in this paper.

Maximum Entropy maximizes entropy i.e. it prefers a uniform distribution when no information is available. Additionally, the approach automatically weights features (predicates). The relationship between neighboring regions is very important in images and Maximum Entropy can account for this in a natural way.

The remainder of the paper is organized as follows. Related work is discussed in section 2. Sections 3 provides a brief description of the features and image vocabulary used while the Maximum Entropy model and its application to image annotation are briefly discussed in 4 Experiments and results are discussed in 5 while Section 6 concludes the paper.

2 Related Work

In image annotation one seeks to annotate an image with its contents. Unlike more traditional object recognition techniques [1, 8, 15, 17] we are not interested in specifying the exact position of each object in the image. Thus, in image annotation, one would attach the label "car" to the image without explicitly specifying its location in the picture. For most retrieval tasks, it is sufficient to do annotation. Object detection systems usually seek to find a specific foreground object, for example, a car or a face. This is usually done by making separate training and test runs for each object. During training positive and negative examples of the particular object in question are presented. However, in the annotation scheme here background objects are also important and we have to handle at least a few hundred different object types at the same time. The model presented here learns all the annotation words at the same time. Object recognition and image annotation are both very challenging tasks.

Recently, a number of models have been proposed for image annotation [2, 4, 7, 9, 12]. Duygulu *et al* [7] described images using a vocabulary of blobs. First, regions are created using a segmentation algorithm like normalized cuts. For each region, features are computed and then blobs are generated by clustering the image features for these regions across images. Each image is generated by using a certain number of these blobs. Their *Translation Model* applies one of the classical statistical machine translation models to translate from the set of blobs forming an image to the set of keywords of an image. Jeon *et al.* [9] instead assumed that this could be viewed as analogous to the cross-lingual retrieval problem and used a *Cross Media Relevance Model* (CMRM) to perform both image annotation and ranked retrieval. They showed that the performance of the model on the same dataset was considerably better than the models proposed by Duygulu *et al.* [7] and Mori *et al.* [11].

The above models use a discrete image vocabulary. A couple of other models use the actual (continuous) features computed over each image region. This tends to give improved results. *Correlation LDA* proposed by Blei and Jordan [4] extends the Latent Dirichlet Allocation (LDA) Model to words and images. This model assumes that a Dirichlet distribution can be used to generate a mixture of latent factors. This mixture of latent factors is then used to generate words and regions. Expectation-Maximization is used to estimate this model. Lavrenko *et al.* proposed the *Continuous Relevance Model* (CRM) to extend the *Cross Media Relevance Model* (CMRM) [9] to directly use continuous valued image features. This approach avoids the clustering stage in in CMRM. They showed that the performance of the model on the same dataset was a lot better than other models proposed.

In this paper, we create a discrete image vocabulary similar to that used in Duygulu *et al* [7] and Jeon *et al.* [9]. The main difference is that the initial regions we use are rectangular and generated by partitioning the image into a grid rather than using a segmentation algorithm. We find that this improves performance (see also [6]). Features are computed over these rectangular regions and then the regions are clustered across images. We call these clusters visterms (visual terms) to acknowledge that they are similar to terms in language.

Berger *et al.*[3] proposed the use of Maximum Entropy approaches for various Natural Language Processing tasks in the mid 1990's and after that many researchers have applied this successfully to a number of other tasks. The Maximum Entropy approach has not been much used in computer vision or imaging applications. In particular, we believe this is the first application of the Maximum Entropy approach to image annotation

3 Visual Representation

An important question is how can one create visterms. In other words, how does one represent every image in the collection using a subset of items from a finite set of items. An intuitive answer to this question is to segment the image into regions, cluster similar regions and then use the clusters as a vocabulary. The hope is that this will produce semantic regions and hence a good vocabulary. In general, image segmentation is a very fragile and erroneous process and so the results are usually not very good.

Barnard and Forsyth[2] and Duygulu *et al.* [7] used general purpose segmentation algorithms like Normalized-cuts[16] to extract regions. In this paper, we use a partition of the image into rectangular grids rather than a segmentation of the image. The annotation algorithm works better when the image is partitioned into a regular grid. than if a segmentation algorithm is used to break up the image into regions (see also [6]). This means that the current state of the art segmentation algorithms are still not good enough to extract regions corresponding to semantic entities. The Maximum Entropy algorithm cannot undo the hard decisions made by the segmentation algorithm. These segmentation algorithms make decisions based on a single image. By using a finer grid, the Maximum Entropy algorithm automatically makes the appropriate associations.

For each segmented region, we compute a feature vector that contains visual information of the region such as color, texture, position and shape. We used K-means to quantize these feature vectors and generate visterms. Each visterm represent a cluster of feature vectors. As in Duygulu et al [7] we arbitrarily choose the value of k. In the future, we need a systematic way of choosing the value.

After the quantization, each image in the training set can now be represented as a set of visterms. Given a new test image , it can be segmented into regions and region features can be computed. For each region, the visterm which is closest to it in cluster space is assigned.

4 Maximum Entropy For Image Annotation

We assume that there is a random process that given an image as an observation generates a label y, an element of a finite set Y. Our goal is to create a stochastic model for this random process. We construct a training dataset by observing the behavior of the random process. The training dataset consists of pairs $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ where x_i represents an image and y_i is a label. If an image has multiple labels, x_i may be part of multiple pairings with other labels in the training dataset. Each image x_i is represented by a vector of visterms. Since we are using rectangular grids, for each position of the cell there is a corresponding visterm.

4.1 Predicate Functions and Constraints

We can extract statistics from the training samples and these observations should match the output of our stochastic model. In Maximum Entropy, any statistic is represented by the expected value of a feature function. To avoid confusion with image features, from now on, we will refer to the feature functions as predicates. Two different types of predicates are used.

- **Unigram Predicate**
  This type of predicate captures the co-occurrence statistics of a visual term and a label. The following is an example unigram predicate that checks the co-occurrence of the label 'tiger' and the visterm v_1 in image x.

$$f_{v_1,tiger}(x, y) = \begin{cases} 1 \text{ if } y = tiger \text{ and } v_1 \in x \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

  If image x contains visual term v_1 and has 'tiger' as a label, then the value of the predicate is 1, otherwise 0. We have unigram predicates for every label and visterm pair that occurs in the training data. Since, we have 125 visual terms and 374 labels, the total number of possible unigram predicates is 46750.

- **Bigram Predicate**
  The bigram predicate captures the co-occurrence statistic of two visterms and a label. This predicate attempts to capture the configuration of the image and the positional relationship between the two visterms is important. Two neighboring visterms are horizontally connected if they are next to each other and their row coordinates are the same. They are vertically connected if they are next to each other and their column coordinates are the same. The following example of a bigram predicate models the co-occurrence of the label 'tiger' and the two horizontally connected visterms v_1 and v_2 in image x.

$$f_{horizontal\_v_1v_2,tiger}(x, y) = \begin{cases} 1 \text{ if } y = tiger \text{ and } x \text{ contains} \\ \quad \text{horizontally connected } v_1, v_2 \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

  If x contains horizontally connected visterms v_1 and v_2 and 'tiger' is a label of x, then the value of the predicate is 1. We also use predicates that captures the occurrence of two vertically connected visterms. In the same way, we can design predicates that use 3 or more visterms or more complicated positional relationships. However, moving to trigrams or even n-grams leads to a large increase in the number of predicates and the number of parameters that must be computed and this requires substantially more computational resources.

The expected value of a predicate with respect to the training data is defined as follow,

$$\tilde{p}(f) = \sum_{x,y} \tilde{p}(x, y) f(x, y) \tag{3}$$

where $\tilde{p}(x, y)$ is a empirical probability distribution that can be easily calculated from the training data. The expected value of the predicate with respect to the output of the stochastic model should be equal to the expected value measured from the training data.

$$\sum_{x,y} \tilde{p}(x, y) f(x, y) = \sum_{x,y} \tilde{p}(x) p(y|x) f(x, y) \tag{4}$$

where $p(y|x)$ is the stochastic model that we want to construct. We call equation 4 a constraint. We have to choose a model that satisfies this constraint for all predicates.

4.2 Parameter Estimation and Image Annotation

In theory, there are an infinite number of models that satisfy the constraints explained in section 4.1. In Maximum Entropy, we choose the model that has maximum conditional entropy

$$H(p) = -\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \tag{5}$$

The constrained optimization problem is to find the p which maximizes $H(p)$ given the constraints in equation 4. Following Berger et al [3] we can do this using Lagrange multipliers. For each predicate, f_i, we introduce a Lagrange multiplier λ_i. We then need to maximize

$$\Lambda(p, \lambda) = H(p) + \sum_i \lambda_i (p(f_i) - \tilde{p}(f_i)) \tag{6}$$

Given fixed λ, the solution may be obtained by maximizing p. This leads to the following equations [3]:

$$p(y|x) = \frac{1}{Z(x)} \exp \left[\sum_{i=1}^{k} \lambda_i f_i(x, y) \right] \tag{7}$$

$$\Psi(\lambda) = -\sum_x \tilde{p}(x) log Z(x) + \sum_i \lambda_i \tilde{p}(f_i) \tag{8}$$

where $Z(x)$ is a normalization factor which is obtained by setting $\sum_y p(y|x) = 1$.

The solution to this problem is obtained by iteratively solving both these equations. A few algorithms have been proposed in the literature including Generalized Iterative Scaling and Improved Iterative Scaling [13]. We use *Limited Memory Variable Metric* method which is very effective for Maximum Entropy parameter estimation [13]. We use Zhang Le's [18] maximum entropy toolkit for the experiments in this paper.

5 Experiment

5.1 Dataset

We use the dataset in Duygulu *et al.*[7] to compare the models. We partition images into 5×5 rectangular grids and for each region, extract a feature vector. The feature vector consists of average LAB color and the output of the Gabor filters. By clustering the feature vectors across the training images, we get 125 visterms.

The dataset has 5,000 images from 50 Corel Stock Photo cds. Each cd includes 100 images on the same topic. 4,500 images are used for training and 500 are used for test. Each image is assigned from 1 to 5 keywords. Overall there are 374 words (see [7]).

5.2 Results

We automatically annotate each test image using the top 5 words and then simulate image retrieval tasks using all possible one word queries. We calculate the mean of the precisions and recalls for each query and also the F-measure by combining the two measures using $F = 1/(\lambda\frac{1}{P} + (1-\lambda)\frac{1}{R})$ where P is the mean precision, R is the mean recall. We set the λ as 0.5.

In this paper, we used the results of the *Translation Model* [7] and the CMRM[9] as our baseline since they also use similar features. The experiment shows that Maximum Entropy using unigram predicates has performance comparable to the CMRM model (both have F-measures of 0.1). While one has better recall, the other has better precision. Both models outperform the classical translation model used by [7]. Using unigram and bigram predicates improves the performance of the Maximum Entropy model. Our belief is that by using predicates which provide even more configuration information, the model's performance can be further improved.

Models which use continuous features (for example [12]) perform even better. Maximum Entropy models can also be used to model such continuous features and future work will involve using such features. The results show that Maximum Entropy models have great potential and also enable us to incorporate arbitrary configuration information in a natural way.

Table 1. Experimental results

Experiment	recall	precision	F-measure
Translation	0.04	0.06	0.05
CMRM	0.09	0.10	0.10
Binary Unigram	0.11	0.08	0.10
Binary Unigram + Binary Bigram	0.12	0.09	0.11

6 Conclusions and Future Work

In this paper we show that Maximum Entropy may be used for the image annotation task and the experimental results show the potential of the approach. Since, we can easily add new types of predicates (this is the one of the nice properties in Maximum Entropy), there is great potential for further improvements in performance. More work on continuous valued predicates, image segmentation techniques and feature extraction methods will also lead to performance improvements.

Fig. 1. Examples: Images in the first row are the top 4 images retrieved by query 'swimmer'. Images in the second row are the top 4 images retrieved by query 'ocean'.

7 Acknowledgments

We thank Kobus Barnard for making their dataset [7] available. We used Zhang Le's publicly available Maximum Entropy Modeling Toolkit [18]. This work was supported in part by the Center for Intelligent Information Retrieval and in part by the National Science Foundation under grant NSF IIS-9909073. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

References

1. S. Agarwal and D. Roth. Learning a Sparse Representation for Object Detection, IN *Proc. ECCV*, pages 113-130, 2002.
2. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107-1135, 2003.
3. A. Berger, S. Pietra and V. Pietra. A Maximum Entropy Approach to Natural Language Processing. In *Computational Linguistics*, pages 39-71, 1996
4. D. Blei, and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th Intl. ACM SIGIR Conf.*, pages 127–134, 2003
5. P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, 19(2):263-311, 1993.
6. P. Carbonetto, N. de Freitas. Why can't Jos read? The problem of learning semantic associations in a robot environment. In *Human Language Technology Conference Workshop on Learning Word Meaning from Non-Linguistic Data*, 2003.
7. P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conf. on Computer Vision*, pages 97-112, 2002.
8. R. Fergus, P. Perona and A Zisserman Object Class Recognition by Unsupervised Scale-Invariant Learning. IN Proc. CVPR'03, vol II pages 264-271, 2003.
9. J. Jeon, V. Lavrenko and R. Manmatha. (2003) Automatic Image Annotation and Retrieval using Cross-Media Relevance Models In *Proceedings of the 26th Intl. ACM SIGIR Conf.*, pages 119–126, 2003

10. V. Lavrenko, M. Choquette, and W. Croft. Cross-lingual relevance models. *Proceedings of the 25th Intl. ACM SIGIR Conf.*, pages 175–182, 2002.
11. Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *MISRM'99 First Intl. Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
12. V. Lavrenko, R. Manmatha and J. Jeon. A Model for Learning the Semantics of Pictures. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, NIPS'03, 2004.
13. Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Workshop on Computational Language Learning*, 2003
14. K. Nigam, J. Lafferty and A. McCallum Using Maximum Entropy for Text Classification In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61-67, 1999
15. H. Schneiderman, T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces and Cars. Proc. IEEE CVPR 2000: 1746-1759
16. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
17. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR'01*, pages 511-518, 2001.
18. Zhang Le. Maximum Entropy Modeling Toolkit for Python and C++. http://www.nlplab.cn/zhangle/