An Inference Network Approach to Image Retrieval

Donald Metzler and R. Manmatha

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

Abstract. Most image retrieval systems only allow a fragment of text or an example image as a query. Most users have more complex information needs that are not easily expressed in either of these forms. This paper proposes a model based on the Inference Network framework from information retrieval that employs a powerful query language that allows structured query operators, term weighting, and the combination of text and images within a query. The model uses non-parametric methods to estimate probabilities within the inference network. Image annotation and retrieval results are reported and compared against other published systems and illustrative structured and weighted query results are given to show the power of the query language. The resulting system both performs well and is robust compared to existing approaches.

1 Introduction

Many existing image retrieval systems retrieve images based on a query image [1]. However, recently a number of methods have been developed that allow images to be retrieved given a text query [2–4]. Such methods require a collection of training images annotated with a set of words describing the image's content. From a user's standpoint, it is generally easier and more intuitive to produce a text query rather than a query image for a certain information need. Therefore, retrieval methods based on textual queries are desirable.

Unfortunately, most retrieval methods that allow textual queries use a rudimentary query language where queries are posed in natural language and terms are implicitly combined with a soft Boolean `AND`. For example, in current models, the query `tiger jet` allows no interpretation other than `tiger AND jet`. What if the user really wants `tiger OR jet`? Such a query is not possible with most existing models. A more richly structured query language would allow such queries to be evaluated.

Finally, an image retrieval system should also allow seamless combination of text and image representations within queries. That is, a user should be able to pose a query that is purely textual, purely based on an image, or some combination of text and image representations.

This paper presents a robust image retrieval model based on the popular Inference Network retrieval framework [5] from information retrieval that successfully combines all of these features. The model allows structured, weighted queries made up of both textual and image representations to be evaluated in a formal, efficient manner.

We first give a brief overview of related work, including a discussion of other image retrieval methods and an overview of the Inference Network retrieval framework in Section 2. Section 3 details our model. We then describe experimental results and show retrieval results from several example queries in Section 4. Finally, we discuss conclusions and future work in Section 5.

2 Related Work

The model proposed in this paper draws heavily from past work in the information and image retrieval fields. This section gives an overview of related work from both fields. A number of models use associations between terms and image regions for image annotation and retrieval. The *Co-occurrence Model* [6] looks at the co-occurrences of annotation words and rectangular image regions to perform image annotation. The *Translation Model* [7] uses a classic machine translation technique to translate from a vocabulary of terms to a vocabulary of blobs. Here, blobs are clusters of feature vectors that can be thought of as representing different "concepts". An unannotated image is represented as a set of blobs and translated into a set of annotation words. The *Cross-Media Relevance Model* [3] (CMRM) also views the task as translation, but borrows ideas from cross-lingual information retrieval [8], and thus allows for both image annotation and retrieval. The *Correspondence LDA Model* [2] (CLDA) allows annotation and retrieval. It is based on Latent Dirichlet Allocation [9] and is a generative model that assumes a low dimensional set of latent factors generate, via a mixture model, both image regions and annotations.

The motivation for the estimation techniques presented here is the *Continuous Relevance Model* [4] (CRMs), which is a continuous version of the CMRM model that performs favorably. Unlike other models that impose a structure on the underlying feature space via the use of blobs, the CRM model uses a non-parametric technique to estimate the joint probability of a query and an image. However, it assumes annotation terms are drawn from a multinomial distribution, which may be a poor assumption. Our estimation technique makes no assumption as to the distribution of annotation terms and thus is fully non-parametric.

The Inference Network retrieval framework is a robust model from the field of information retrieval [5] based on the formalism of Bayesian networks [10]. Some strong points of the model are that it allows structured, weighted queries to be evaluated, multiple document representations, and efficient inference. One particular instantiation of the inference network framework is the InQuery retrieval system [11] that once powered Infoseek and currently powers the THOMAS search engine used by the Library of Congress. Additionally, inference networks

for multimedia retrieval in extensible databases have been explored [12]. Experiments have shown that intelligently constructed structured queries can translate into improved retrieval performance. Therefore, the Inference Network framework is a good fit for image retrieval.

3 Image Inference Network Model

Suppose we are given a collection of annotated training images $\mathcal{T}$. Each image $I \in \mathcal{T}$ has a fixed set of words associated with it (annotation) that describe the image contents. These words are encoded in a vector tf_I that contains the number of times each word occurs in I's annotation. We also assume that I has been automatically segmented into regions. Note that each image may be segmented into a different number of regions. A fixed set of d features is extracted from each of these regions. Therefore, a d dimensional feature vector r_i is associated with each region of I. Finally, each feature vector r_i extracted from I is assumed to be associated with each word in I's annotation. Therefore, tf_I is assumed to describe the contents of each r_i in I. Images in the test set are represented similarly, except they lack annotations. Given a set of unseen test images, the image retrieval task is to return a list of images ranked by how well each matches a user's information need (query).

3.1 Model

Our underlying retrieval model is based on the Inference Network framework [5]. Figure 1 (left) shows the layout of the network under consideration. The node J is continuous-valued and represents the event an image is described by a collection of feature vectors. The q_{w_j} nodes are binary and represent the event that word w_j is observed. Next, the q_{r_k} nodes are binary and correspond to the event that image representation (feature vector) r_k is observed. Finally, the q_{op} and I nodes represent query operator nodes. I is simply a special query operator that combines all pertinent evidence from the network into a single belief value representing the user's information need. These nodes combine evidence about word and image representation nodes in a structured manner and allow efficient marginalization over their parents [13]. Therefore, to perform ranked image retrieval, for each image X in the test collection we set $J = X$ as our observation, run belief propagation, and rank documents via $P(I|X)$, the probability the user's information need is satisfied given that we observe image X.

Figure 1 (right) illustrates an example instantiation of the network for the query `#OR(#AND(tiger grass) <tiger.jpg>)`. This query seeks an image with both `tigers` AND `grass` OR an image that is similar to `tiger.jpg`. The image of the tiger appearing at the top is the image currently being scored. The other nodes, including the cropped tiger image, are dynamically created based on the structure and content of the query. Given estimates for all the conditional probabilities within the network, inference is done, and the document is scored based on the belief associated with the `#OR` node.

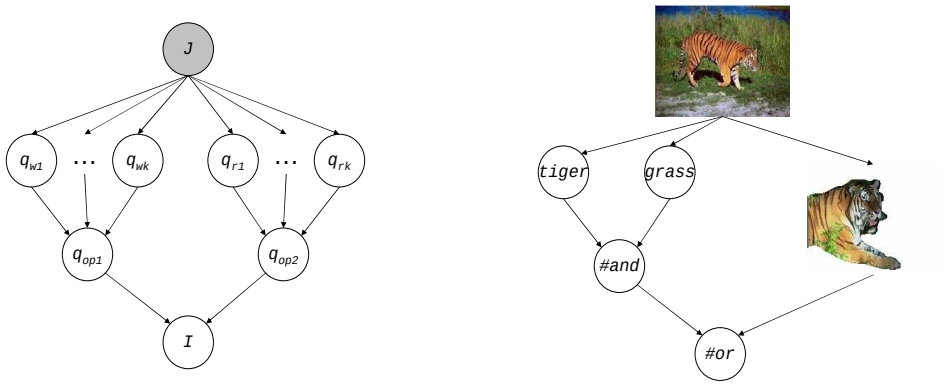

Fig. 1. Inference network layout (left) and example network instantiation (right)

Now that the inference network has been set up, we must specify how to compute $P(q_w|J)$, $P(q_r|J)$, and the probabilities at the query operator nodes. Although we present specific methods for estimating the probabilities within the network, the underlying model is robust and allows any other imaginable estimation techniques to be used.

3.2 Image representation node estimation

To estimate the probability of observing an image representation r given an image J we use a density estimation technique. The resulting estimate gives higher probabilities to representations that are "near" (or similar to), on average, the feature vectors that represent image J. Thus, the following estimate is used:

$$P(q_r|J) = \frac{1}{|r_J|} \sum_{r_i \in J} \mathcal{N}(q_r; r_i, \Sigma)$$

where $|r_J|$ is the number of feature vectors associated with J and,

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

is a multivariate Gaussian kernel. Here, Σ is assumed to be diagonal and is the empirical variance with respect to the training set feature vectors. Note that any appropriate kernel function can be used in place of $\mathcal{N}$.

3.3 Term node estimation

Estimating the likelihood a term is observed given an image, $P(q_w|J)$, is a more difficult task since test images are not labeled with annotations. Inverting the

probability by Bayes' rule and assuming feature vectors are conditionally independent given q_w we see that

$$P(q_w|J) \propto P(q_w)P(J|q_w) = P(q_w) \prod_{r_i \in J} P(r_i|q_w)$$

where $P(q_w) = \frac{n_{q_w}}{n_{tot}}$, n_{q_w} is the number of feature vectors in the training set q_w is associated with and n_{tot} is the total number of feature vectors in the training set. To compute $P(r_i|q_w)$ we again use density estimation to estimate the density of feature vectors among the training set that are associated with term q_w. Then,

$$P(r_i|q_w) = \frac{1}{n_{q_w}} \sum_{\substack{I \in \mathcal{T} \\ tf_I(q_w)>0}} \sum_{g_k \in I} \mathcal{N}(r_i; g_k, \Sigma)$$

where $tf_I(q_w)$ indicates the number of times that q_w appears in image I's annotation and $\mathcal{N}$ is defined as above. Finally, it should be noted that when Σ is estimated from the data our model does not require hand tuning any parameters.

3.4 Regularized term node estimates

As we will show in Section 4, the term node probability estimates used here result in good annotation performance. This indicates that our model estimates probabilities well for single terms. However, when combining probabilities from multiple term nodes we hypothesize that it is often the case that the *ordering* of the term likelihoods captures more important information than their actual values. That is, the fact that $tiger = \arg\max_{q_w} P(q_w|J_1)$ is more important than the fact that $P(tiger|J_1) = 0.99$. Thus, we explore regularizing the term node probabilities for each image so they vary more uniformly and are based on the ordering of the term likelihoods.

Assuming the term likelihood ordering is important and that for an image: 1) a few terms are very relevant (correctly annotation terms), 2) a medium number of terms are somewhat relevant (terms closely related to the annotation terms), and 3) a large number of terms are irrelevant (all the rest). Following these assumptions we fit the term probability estimates to a Zipfian distribution. The result is a distribution where a large probability mass is given to relevant terms, and a smaller mass to the less relevant terms. We assume that terms are relevant based on their *likelihood rank*, which is defined as the rank of term w in a sorted (descending) list of terms according to $P(q_w|J)$. Therefore, the most likely term is given rank 1. For an image J we regularize the corresponding term probabilities according to $\hat{P}(q_w|J) = Z^{-1}\frac{1}{R_{q_w,J}}$ where $R_{q_w,J}$ is the likelihood rank of term w for image J and Z^{-1} normalizes the distribution.

3.5 Query Operators

Query operators allow us to efficiently combine beliefs about query word nodes and image representation nodes in a structure (logical) and/or weighted manner. They are defined in such a way as to allow efficient marginalization over

their parent nodes [14], which results in fast inference within the network. The following list represents a subset of the structured query operators available in the InQuery system that have been implemented in our system: `#AND`, `#WAND`, `#OR`, `#NOT`, `#SUM`, and `#WSUM`. To compute $P(q_{op} = true|J)$, the belief at query node q_{op}, we use the following:

$$P_{\#AND}(q_{op}|J) = \prod_i p_i \qquad P_{\#WAND}(q_{op}|J) = \prod_i p_i^{\frac{w_i}{W}}$$

$$P_{\#SUM}(q_{op}|J) = \frac{1}{n}\sum_i p_i \qquad P_{\#WSUM}(q_{op}|J) = \frac{1}{W}\sum_i w_i p_i$$

$$P_{\#OR}(q_{op}|J) = 1 - \prod_i (1 - p_i) \qquad P_{\#NOT}(q_{op}|J) = 1 - p_1$$

where node q_{op} has n parents $\pi_1, \ldots, \pi_n$, $p_i = P(\pi_i|J)$, and $W = \sum_i w_i$. See [11, 14, 13] for a derivation of these expressions, an explanation of these operators, and details on other possible query operators.

4 Results

We tested our system using the Corel data set that consists of 5000 annotated images. Each image is annotated with 1-5 words. The number of distinct words appearing in annotations is 374. The set is split into 4500 training images and 500 test images. Each image is segmented using normalized cuts into 1-10 regions. A standard set of 36 features based on color and texture is extracted from each image region. See [7] for more details on the features used. We compare the results of our system with those reported from the Translation, CMRM and CRM models that used the same segmentation and image features. Throughout the results, `InfNet-reg` refers to the model with regularized versions of the term probabilities and `InfNet` refers to it with unregularized probabilities.

4.1 Image Annotation

The first task we evaluate is image annotation. Image annotation results allow us to compare how well different methods estimate term probabilities. The goal is to annotate unseen test images with the 5 words that best describe the image. Our system annotates these words based on the 5 terms with the highest likelihood rank for each image. Mean per-word recall and precision are calculated, where recall is the number of images correctly annotated with a word divided by the number of images that contain that word in the human annotation, and precision is the number of images correctly annotated with a word divided by the total number of images annotated with that word in the test set. These metrics are computed for every word and the mean over all words and are reported in Table 1. As the table shows, our system achieves very good performance on the annotation task. It outperforms CRMs both in terms of mean word precision and recall, with the mean per-word recall showing a 26.3% improvement over the CRM model.

Table 1. Annotation results

Models	Translation	CMRM	CRM	InfNet
# words w/ recall > 0	49	66	107	112
Results on full vocabulary				
Mean per-word recall	0.04	0.09	0.19	0.24
Mean per-word precision	0.06	0.10	0.16	0.17

4.2 Image Retrieval

For the retrieval task we create all 1-, 2-, 3- word queries that are relevant to at least 2 test images. An image is assumed to be relevant if and only if its annotation contains every word in the query. Then, for a query $Q = q_1, q_2, \ldots, q_L$ we form and evaluate a query of the form `#and(`$q_1, \ldots, q_L$`)`. We use the standard information retrieval metrics of mean average precision (MAP) and precision at 5 ranked documents to evaluate our system. Table 2 reports the results.

Table 2. Retrieval results and comparison

Query length	1 word	2 word	3 word
Number of queries	179	386	178
Relevant images	1675	1647	542
Precision after 5 retrieved images			
CMRM	0.1989	0.1306	0.1494
CRM	0.2480	0.1902	0.1888
InfNet	0.2525	0.1672	0.1727
InfNet-reg	0.2547	0.1964	0.2170
Mean Average Precision			
CMRM	0.1697	0.1642	0.2030
CRM	0.2353	0.2534	0.3152
InfNet	0.2484	0.2155	0.2478
InfNet-reg	0.2633	0.2649	0.3238

The regularized probabilities result in much better retrieval performance over the unregularized probabilities. Using these probabilities, our system achieves better performance than both the CMRM and CRM models on all 3 query sets for both the MAP and precision after 5 retrieved documents metric.

Figure 2 gives illustrative examples of the top 4 ranked documents using the regularized estimates for several structured queries. The first query of Figure 2, `#or(swimmers jet)`, results in images of both swimmers and jets being returned. The next query shows that a standard query gives good retrieval results. The next two queries demonstrate how term weighting can affect the retrieval results. Finally, the last query shows an example query that mixes text and image representations, with the bird image being part of the query. These results show that the structured operators can be used to form rich, powerful queries that other approaches are not capable of.

Fig. 2. Example structured query results

5 Conclusions and Future Work

We have presented an image retrieval system based on the inference network framework from information retrieval. The resulting model allows rich queries with structured operators and term weights to be evaluated for combinations of terms and images. We also presented novel non-parametric methods for estimating the probability of a term given an image and a method for regularizing the probabilities. Our system performs well compared to other published models under standard experimental evaluation.

There are a number of things that can be done as part of future work. First, better estimates for $P(q_r|J)$ are needed. The method described in this paper was used for simplicity. Second, our system must be tested using different segmentation and better features to allow comparison against other published results. Third, more rigorous experiments should be done using the structured and weighted query operators to show empirically what affect they have on overall performance. Finally, it would be interesting to explore a model that combines the current system with a document retrieval system to allow for full text and image search in a combined model.

Acknowledgments

We thank Kobus Barnard for making the Corel dataset available at http://vision.cs.arizona.edu/kobus/research/data/eccv_2002/.

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the National Science Foundation under grant number IIS-9909073 and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

1. Flickner, M., Sawhney, H.S., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system. IEEE Computer Magazine **28** (Sept. 1995) 23–30
2. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference. (2003) 127–134
3. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference. (2003) 119–126
4. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems. (2003)
5. Turtle, H., Croft, W.B.: Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems (TOIS) **9** (1991) 187–222
6. Mori, Y., Takahashi, H., , Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management. (1999)
7. Duygulu, P., Barnard, K., de Fretias, N., Forsyth, D.: Object recognition as machine translation: Learning a leicon for a fixed image vocabulary. In: Seventh European Conference on Computer Vision. (2002) 97–112
8. Lavrenko, V., Choquette, M., Croft, W.: Cross- lingual relevance models. In: Proceedings of the 25th annual International ACM-SIGIR conference. (2002)
9. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
10. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc. (1988)
11. Callan, J.P., Croft, W.B., Harding, S.M.: The INQUERY retrieval system. In: Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications. (1992) 78–83
12. de Vries, A.: Mirror: Multimedia query processing in extensible databases. In: Proceedings of the fourteenth Twente workshop on language technology (TWLT14). (1998) 37–48
13. Turtle, H.R.: Inference Networks for Document Retrieval. PhD thesis, University of Massachusetts (1991)
14. Greiff, W.R., Croft, W.B., Turtle, H.: PIC matrices: a computationally tractable class of probabilistic query operators. ACM Transactions on Information Systems **17** (1999) 367–405