Understanding and Predicting Importance for Abstract Images

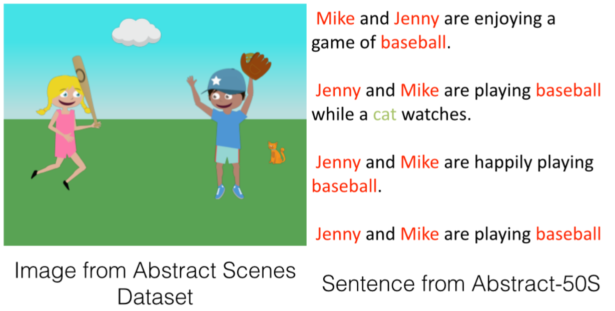
Shanmukha Ramakrishna Vedantam
Virginia Tech
vrama91@vt.edu

Abstract

What makes an object important in an image? Is it possible to predict given an image, how important a particular object is? An understanding of importance can help tune features that better predict salient semantic (important) aspects of a scene. Inspired by previous work, I use sentences describing images as a proxy for high level semantic importance. Given such sentences and abstract images, I attempt to understand importance by picking rich semantic features (gaze, expression, pose) and evaluating performance at predicting importance. Results show that the semantic features give a significant improvement in performance over simple location based features. Interestingly, it is found that after humans, animal categories are judged to be the most important. This might point towards evolutionary factors at play.

1. Introduction

The problem of understanding the importance of objects in an image is central to semantic scene understanding. What is it that makes one object more important than another object in the image? Answering this question reliably has applications applications to Image Search - where Images in which the object of interest is more important could be ranked higher and shown to a user. Apart from this, understanding importance is also useful for doing better text/description generation from images. It is also interesting to explore category level importance and check for any evolutionary biases in humans in terms of humans or animals being judged more important [1, 4]. To answer what makes an object important, one might think of using location location or low level saliency features. However, since importance is essentially a semantic concept and due to recent advances in object detection, works have started focussing on importance at an object-attribute-scene level [1]. However there might be other factors relating to the gaze of the objects in the scene, their expression, gender, pose etc. that might be significant. These are not generic attributes. It is interesting to think of an object making a person in an image happy or resulting in a certain pose of the person and thus becoming more important. While a previous work [1] has used factors related to composition, semantics and context for importance,the semantic and context features are typically global. In contrast to this work, I work with a larger feature set on a dataset of abstract images [10], where I explore semantics on a more image specific level. Also, I reason about importance of all the objects in the image jointly using a more expressive Conditional Random Field model. Other works have explored importance as the ordering in which objects are mentioned in tag lists [2]. Unlike this work, I consider full sentences for importance. All objects that have been mentioned in the sentence are equally important. The interplay between eye fixations, image features and human descriptions has also been studied in [9]. While they study gaze movements of subjects looking at the image, I reason about gaze of the objects within the image.



Mike and Jenny are enjoying a game of baseball.

Jenny and Mike are playing baseball while a cat watches.

Jenny and Mike are happily playing baseball.

Jenny and Mike are playing baseball

Image from Abstract Scenes Dataset
Sentence from Abstract-50S

Figure 1. Words indicated in red (right) are the objects in the image which got mentioned in human sentences.

The rich semantic annotation in Abstract Images makes this study feasible. A fine grained study of importance from the perspective of Human Agreement has also been made possible with a dataset of 50 sentences for these Abstract Images.

2. Approach

I now describe my approach in the following sections. Section 2.1 explains my methodology for calculating ground truth importance and computation of human agreement. Section 2.2 explains my model and its parameters. Section 2.3 explains the node features. Section 2.4 explains the edge features.

2.1. Calculating Ground Truth Importance

The ground truth importance is computed as a soft occurrence score. This is done since the object labels in the dataset might not be mentioned verbatim in human sentences. Thus it is important to account for cases where for instance,"ketchup" is mentioned as "food" and so on. This is the problem of synecdoche. Apart from this, there also exist issues of synonymy and verb modifiers. A word net based lexical similarity measure [8] is used to compute the soft occurrence in a sentence, similar to [1]. That is, given an object label in the image, the similarity to all the words in the sentence is computed. The maximum value of this similarity is set as the soft occurrence score for a particular sentence. Given more sentences, the mean occurrence across multiple sentences is taken. Experimentally, in my dataset this mean is computed across 48 sentences.

Given soft importance scores per sentence, a threshold is set based on optimizing for human agreement at 48 sentences. This is done by calculating importance across two random splits of data. The importance of one split is treated as the prediction and the other split is treated as the ground truth. Class normalized accuracy is computed.

2.2. CRF model

Inference on a conditional random field model, instantiated on each particular image gives the importance estimates. A node is placed in the CRF for each object in the image, similar to [3]. Two alternative topologies are considered: a fully connected model and a category connected model. The fully connected CRF model incorporates an edge between every pair of nodes. Thus it tries to model the importance of each object in the image by considering the influence of every other object. The category level connected model builds edges only between nodes which belong to the same category, eg. animals in the Abstract scenes dataset. There are 8 such categories in the dataset. The idea here is to capture pairwise dependencies in within a category. Given Y and X, the CRF model is parameterized as follows:

$$logP(Y|X) = \sum_{i=1}^{N} (\alpha * \psi(y_i; X) + \beta * \phi(y_i; X)) + \sum_{i,j} \lambda * \delta(y_i, y_j) + logZ \quad (1)$$

where α, β, λ are the parameters of the CRF. $logZ$ is the log-partition function of the CRF. N is the number of objects in the image. ψ are the semantic node potentials. ϕ are the location and occurence node potentials and δ are the edge potentials of the CRF.

2.3. Node Potentials

Two node potentials are computed for each node in the CRF. The first is an SVM trained on semantic features. Semantic features are a histogram of the number of objects in the image in a certain pose (out of 7 possible poses) concatenated with a similar count for expression (out of 5 possible expressions) , gaze features and category features. The gaze feature for an object is the number of other objects in the image have the object in their gaze field. The gaze field is defined presently by drawing a rectangular box from the eyeline of the object in question. More sophisticated gaze models could be considered. The category features are an indicator vector of what category the object belongs to. Using these features, a linear SVM is trained to give semantic potentials as confidence. The second node potential is computed as follows. A location potential is defined for each node in the CRF which takes the location in x, y and depth and the object instance as the input. There are 58 object instances in the dataset. Thus this is a feature of length 61. Similar to the semantic features, a linear SVM is trained with the location features and the prediction confidences are used as the location potentials.

2.4. Edge Potentials

I experimented by plugging in three variations of edge potentials into the CRF. Two of them are **Text Potentials** which are derived from text co-occurrence statistics. The first kind of text potential is computed as follows. Firstly, Importance is estimated for all objects that occur in this in a large set of 5000 images. For each pair of objects the set of common images they co-occur in is found. A correlation edge potential is then defined as the correlation between importances of the two objects in this set. Intuitively, this captures how much the importance of one object tells us about the other. This potential places no reward or penalty for disagreement. It only operates on agreement. Thus $\delta(i, j) = 0$ when $i = j$. These potentials are referred to as Correlation Potentials (CP) in the rest of the report.

The second kind of text potential considers importance as a probability. For all the four possible states of two binary variables, edge potentials are estimated as $\delta(s_i, s_j) = \sum_k p_1^k(s_i) * p_2^k(s_j)$ where $s_i = \{0, 1\}$ and p_1^k is the probability of the first object being important in image k. The sum then goes over all the images in which object 1 and 2 co-occur, just like the previous text potential. These are referred to as Probability Potentials (PP) in the rest of the report.

The third kind of edge potentials are **Image Potentials** based on location of the two objects in the Image. These potentials capture any dependence of importance to the distance between two objects. The distance is defined on a feature space with location and object category as features and set as the edge potential for rewarding similar states of objects. The idea is to encode notions of similar objects in vicinity being similar or dissimilar in importance.

2.5. Parameter Selection and Inference

Parameter selection is done via cross validation on a held out set. A grid search is done over the three parameters (α, β, λ). Inference is done via Iterated Conditional Modes.

3. Experiments and Results

Dataset: The dataset used for learning node potentials and testing the model is Abstract-50S which contains 500 images with 50 associated sentences. The images are permuted at random and the first 300 images are used as the train set, 100 are used as the val set and 100 images are used as the test set. The Text Potentials are learnt from a subset of 5000 images of the Abstract Scenes Dataset [10].

Implementation: Most of the implementation has been done on MATLAB. For inference on CRF (ICM) the Undirected Graphical Model (UGM) - Toolbox is used [7]. The code for synset based word similarity was written in python using the Natural Language Toolkit (NLTK) [5].

Human Agreement: There are two design choices for calculating ground truth importance. The first is to answer how many sentences are needed for computing importance reliably. Secondly, since importance estimation is essentially a soft occurrence score, a threshold is needed to convert it into binary labels.The choice of this threshold is the second design choice. Both are solved by optimizing for Human Agreement. Importance values were thresholded at varying thresholds and class normalized accuracy was computed across multiple train test splits for split size of 24 sentences (Fig. 2). Looking at the plot, the threshold was set at 0.5 since it gives a high human agreement and it seems a natural threshold to choose. Fig. 2 also shows that the human agreement value reaches as high as 96 %. Therefore with 48 sentences per Image, importance is fairly well defined. To further understand the growth of human agreement with respect to sentences, a second test was done which varied the size of the splits. The results are shown in Fig. 3. It can be seen that even when only 1 sentence is present per split, there is an agreement of 82 % on importance.

Baselines: Four baselines are considered. The first is random, where an object is always predicted as unimportant. This gives an accuracy of 53 %. The second baseline predicts mike and jenny to be always important. This gives an accuracy of 81.13 %. Please note that this is unnormalized accuracy. The third baseline considers average category importance across the dataset. Given an image, it predicts an object to be important if the category level importance score is greater than 0.5.This gives an accuracy of 83.6 %. Fig. 4 shows category level importance scores. Interestingly, the most important classes turn out to be Mike, Jenny and Animal. This is in agreement with the result of [1]. The final baseline is an instance level importance, which calls an object as important if on average it is important more often than not across the dataset. This gives an accuracy of 87.4 %.

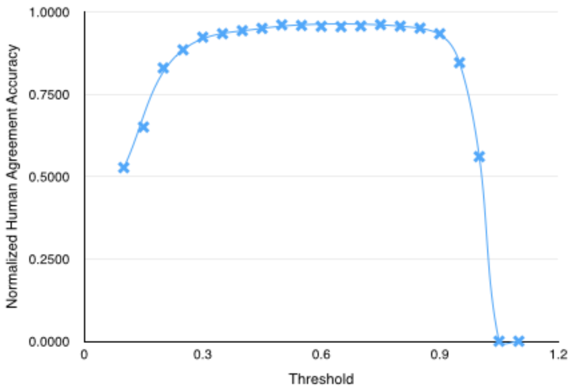


Figure 2. Human Agreement on importance for Abstract - 50S at different thresholds

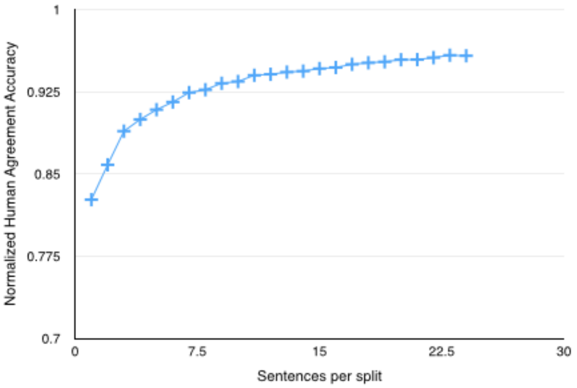


Figure 3. Human Agreement with more sentences on Abstract 50S

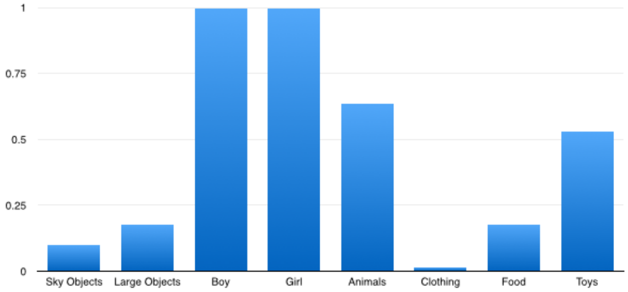


Figure 4. Category Level Importance Values for Abstract-50S

Node Potential Predictions: Next, an experiment was done to check the accuracy obtained on importance classification using only node features. This was done for both

Method	Test Set Accuracy	Val Set Accuracy
FCPP.	78.5	86.9
FCCP	78.0	78.5
PCPP	**85.5**	87.4
PCCP	83.9	87.9

Table 1. Results on Test Set for Abstract-50S importance prediction. FCPP - Fully Connected Probability Potential, FCCP-Fully Connected Correlation Potential, PCPP - Partially Connected Probability Potential, PCCP - Partially Connected Correlation Potential

semantic as well as location node features described in the previous section. The semantic features give an accuracy of 85.2 % on the test set. The location features give an accuracy of 63.4% on the test set. Clearly, the semantic features outperform the location features. However, since the information captured by them is complementary, it still makes sense to keep both the potentials for the final CRF model.

Full CRF Predictions: Table 1 shows the results for importance prediction using Text Based edge potentials. All results are shown after cross val on the Val Set. It can be seen that in general, the category connected CRF topology performs better than the fully connected CRF, despite the latter being the more expressive model. This might be because more data might be needed to reliably train the CRF which operates in a larger model space. This is visible from the fact that we see that for the Probability based potentials, the performance comes down rapidly from the Val set to the Test set. In general we find that the best performing model is the Partly Connected Probability Potential model at 85.5%. From the best node potential performance, this is a jump of 0.3 %, which is not very significant. None of the full CRF predictions are able to beat the instance level importance prediction baseline (87.4 %). For the Party Connected Correlation Potentials case, the Image Based edge Potentials give 67.2 % accuracy. The results from Image Based edge Potentials are not competitive with the Text Based Edge Potentials.

4. Conclusion and Future Work

This project was an attempt to understand and predict importance for abstract images. While there are previous approaches to predict importance, this work explores a more semantic feature set available from a richly (trivially) annotated abstract scenes dataset. From the perspective of understanding importance, the key results are in the fact that animal, human and toy classes are the most important classes. Further, there is significant agreement between humans about importance even when a fairly small number of sentences per image are present. In terms of prediction, the semantic features are shown to perform better than location potentials. However, these methods are up against very strong baselines which have proven difficult to beat. My approaches presently only manage to beat three out of four baselines. The CRF model improves over the node potential prediction performance weakly. In particular, the Partly Connected model outperforms the Fully Connected model.

In terms of Future Work, I would like to explore some better topologies for CRF node connections which strike a balance between the present FC and PC models. Further, I would like to apply this approach to learn importance models from a larger training dataset. This is a practical direction, since Fig. 3 shows that the Human Agreement is significant even if 8 sentences are collected per image and is above present baseline performance. Further, I would like to think about regressing to importance values. Since regression is more fine grained, there are chances of the baselines not performing as well as they currently do. Another interesting direction of work is to explicitly incorporate features that answer the “Why” questions about an image. For instance, the answer to “Why is jenny happy?” might be an important object in the image.

References

[1] A. C. Berg, T. L. Berg, H. D. III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569. IEEE, 2012.

[2] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3:3):1–15, Mar 2008.

[3] D. F. Fouhey and C. Zitnick. Predicting object dynamics in scenes. In *CVPR*, 2014.

[4] F. F. Li, R. Van Rullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academies of Sciences of the United States of America*, 99(14):9596–9601, 2002.

[5] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP ’02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[6] H. S. Park, E. Jain, and Y. Sheikh. Predicting primary gaze behavior using social saliency fields. December 2013.

[7] M. Schmidt. http://www.di.ens.fr/ mschmidt/software/ugm.html. 2007.

[8] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[9] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *CVPR*, pages 739–746. IEEE, 2013.

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