# Image Specificity
# (Supplementary material)


Mainak Jas
Aalto University
mainak.jas@aalto.fi

Devi Parikh
Virginia Tech
parikh@vt.edu


## 1. Scatter plots for correlations

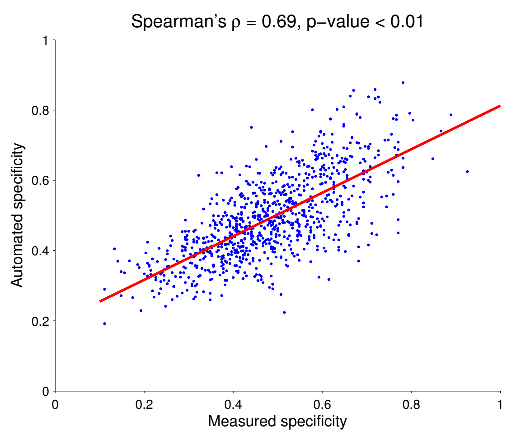


Figure 1. Correlation between human-measured specificity and automated specificity for the MEM-5S dataset.

In the main paper, we described how automated specificity correlated with human-measured specificity. Figure 1 further illustrates this using a scatter plot. We also studied how various image properties correlated with specificity. In Figure 2, we illustrate these correlations via scatter plots.

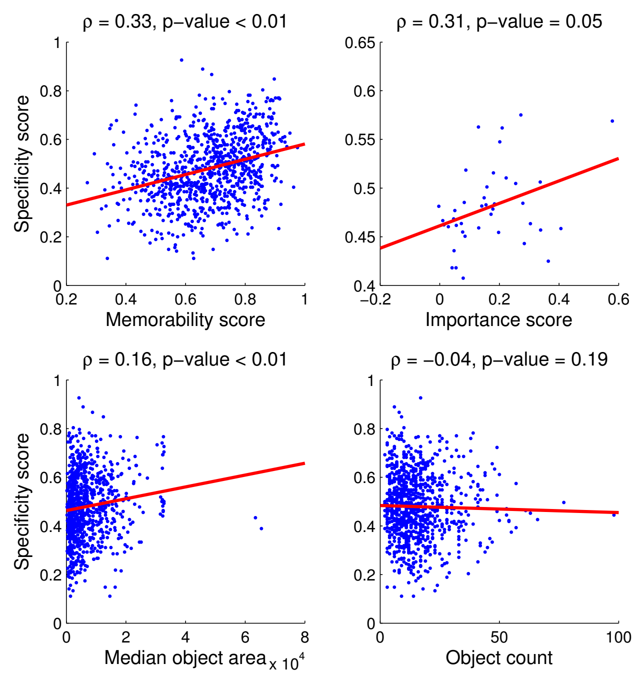


Figure 2. What makes an image specific? Memorable images, images with large objects and important object categories tend to be more specific. Number of annotated objects in an image does not correlate with specificity. Results are on the MEM-5S dataset.

## 2. Predicting specificity

As we have shown, certain image-level objects and attributes make some images more specific than others. This means that specificity may be predictable using image features alone.

To test this, a $\nu$-SVR with an RBF kernel is trained on a randomly chosen subset of images represented by their DECAF-6 features [2] in the MEM-5S and PASCAL-50S datasets. In the ABSTRACT-50S dataset, the image features are a concatenation of object occurrence, their absolute position, depth, flip angle, object co-occurrence, and clip art category [6]. For prediction, 188 images are set aside in the MEM-5S dataset, 200 images in the PASCAL-50S dataset, and 100 images in the ABSTRACT-50S dataset. Figure 3 shows that as the number of images used for training increases, the correlation of the predicted specificity with the ground truth automated specificity increases. We see that specificity can indeed be predicted from just image content better than chance. The use of semantic features (*e.g.* occurence of objects) as opposed to low-level features (*e.g.* DECAF-6) in the ABSTRACT-50S dataset seem to make it easier to predict specificity for that dataset as compared to the MEM-5S and PASCAL-50S datasets. Note that here we are directly predicting automated specificity whereas in the main paper, we focused

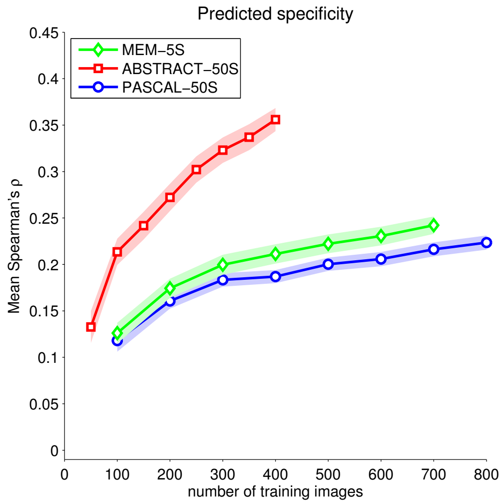


Figure 3. Spearman's rank correlation between predicted and automated specificity for increasing number of training images (averaged across 50 random runs). Automated specificity (Section 3.1.2 in main paper) uses 5, 48 and 50 sentences per image for the three datasets, MEM-5S, ABSTRACT-50S and PASCAL-50S to estimate the specificity of the image. Predicted specificity (Section 2) uses only image features to predict the specificity. Different datasets have different number of images in them, hence they stop at different points on the x-axis. Higher correlation is better. The error bars represented by shaded colors show the standard error of the mean (SEM).

on predicting the two parameters of the Logistic Regression model. The latter is directly relevant to the image search application on which we demonstrated the benefit of specificity.

## 3. Detailed explanation of automated specificity computation

In Figure 4, we visually illustrate the equations and notations used to automatically compute the similarity between two sentences (described in Section 3.1.2 in the main paper). To measure specificity automatically given the $N$ descriptions for image $i$, we first tokenize the sentences and only retain words of length three or more. This ensured that semantically irrelevant words, such as 'a', 'of', *etc.*, were not taken into account in the similarity computation (a standard stop word list could also be used instead). We identified the synsets (sets of synonyms that share a common meaning) to which each (tokenized) word belongs using the Natural Language Toolkit [1]. Words with multiple meanings can belong to more than one synset. Let $Y_{au} = \{y_{au}\}$ be the set of synsets associated with the $u$-th word from sentence $s_a$.

Every word in both sentences contributes to the automatically computed similarity $sim_{\text{auto}}(s_a, s_b)$ between a pair

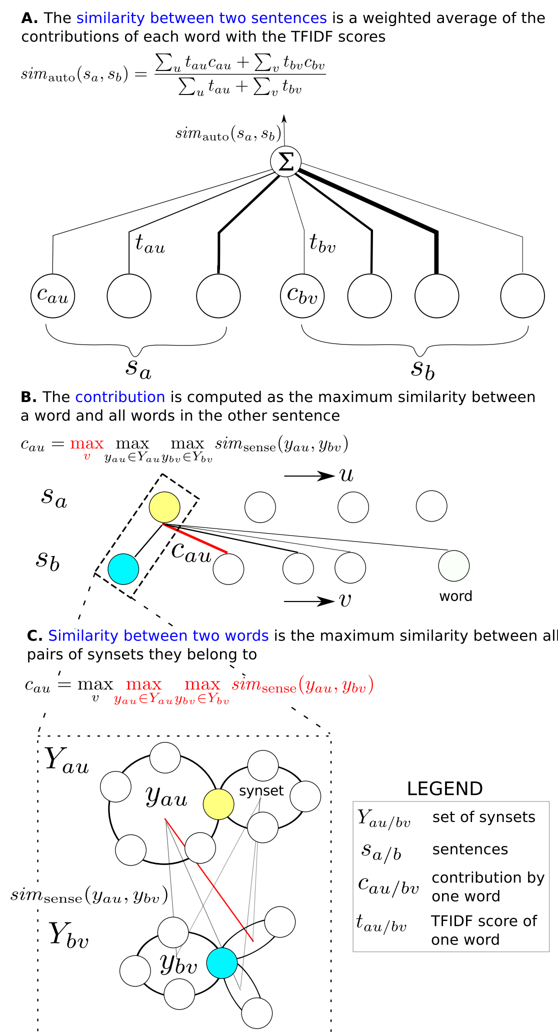


Figure 4. Illustration of our approach to compute automated sentence similarity.

of sentences $s_a$ and $s_b$. The contribution of the $u$-th word from sentence $s_a$ to the similarity is $c_{au}$. This contribution is computed as the maximum similarity between this word, and all words in sentence $s_b$ (indexed by $v$) (Figure 4B). The similarity between two words is the maximum similarity between all pairs of synsets (or senses) to which the two words have been assigned (Figure 4C). We take the maximum because a word is usually used in only one of its senses. Concretely,

$$c_{au} = \max_{v} \max_{y_{au} \in Y_{au}} \max_{y_{bv} \in Y_{bv}} sim_{\text{sense}}(y_{au}, y_{bv}) \qquad (1)$$

The similarity between senses $sim_{\text{sense}}(y_{au}, y_{bv})$ is the shortest path similarity between the two senses on WordNet [4]. We can similarly define $c_{bv}$ to be the contribution of $v$-th word from sentence $s_b$ to the similarity $sim_{\text{auto}}(s_a, s_b)$ between sentences $s_a$ and $s_b$.

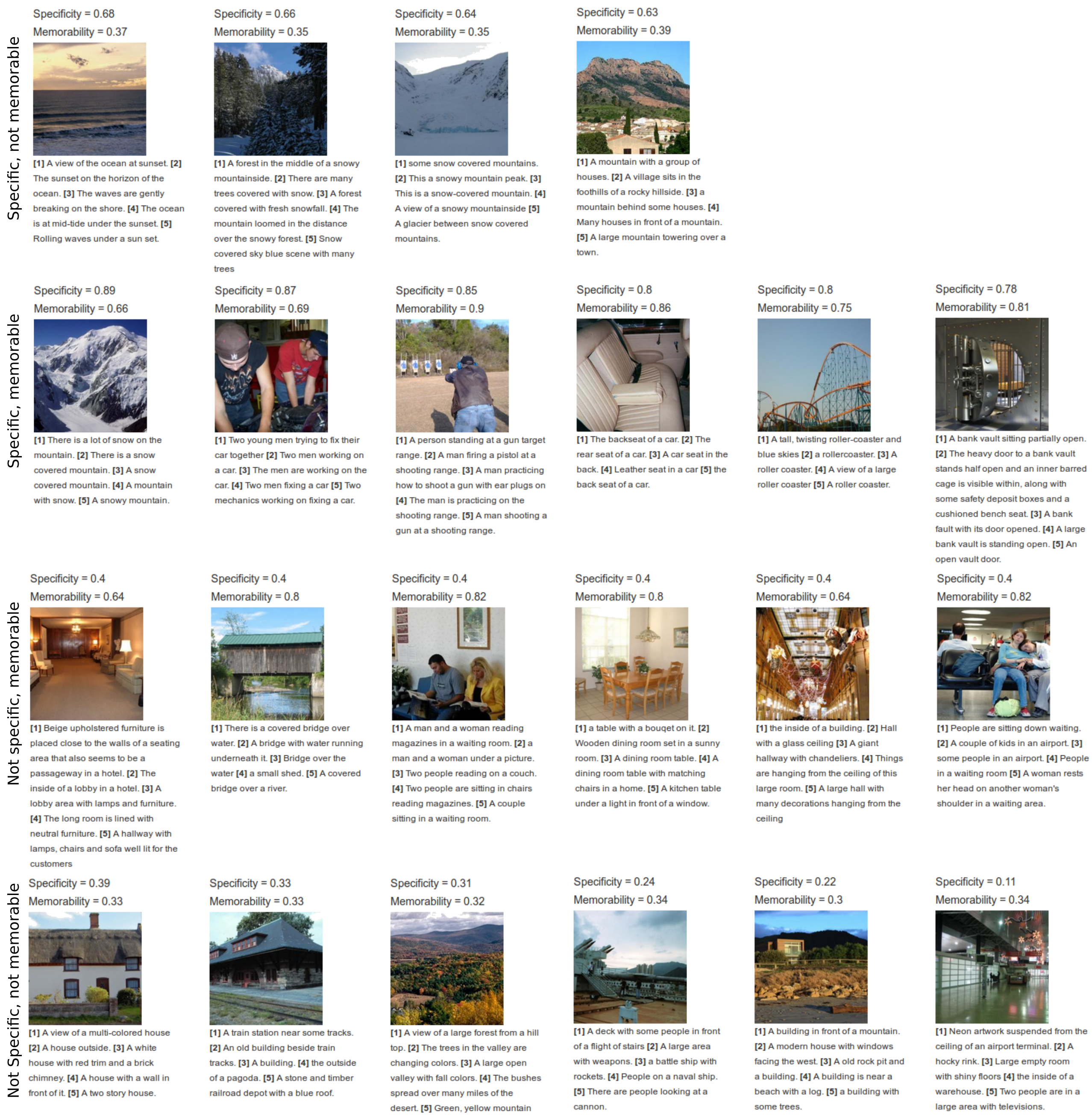


Figure 5. Examples illustrating the similarity and distinctions between image memorability [3] and image specificity.

The similarity between the two sentences is defined as the average contribution of all words in both sentences, weighted by the importance of each word (Figure 4A). Let the importance of the $u$-th word from sentence $s_a$ be $t_{au}$. This importance is computed using term frequency-inverse document frequency (TF-IDF) using the scikit-learn software package [5]. Words that are rare in the corpus but occur frequently in a sentence contribute more to the similarity of that sentence with other sentences. So we have

$$sim_{\text{auto}}(s_a, s_b) = \frac{\sum_u t_{au} c_{au} + \sum_v t_{bv} c_{bv}}{\sum_u t_{au} + \sum_v t_{bv}} \quad (2)$$

The denominator in Equation 2 ensures that the similarity between two sentences is independent of sentence-length and is always between 0 and 1.

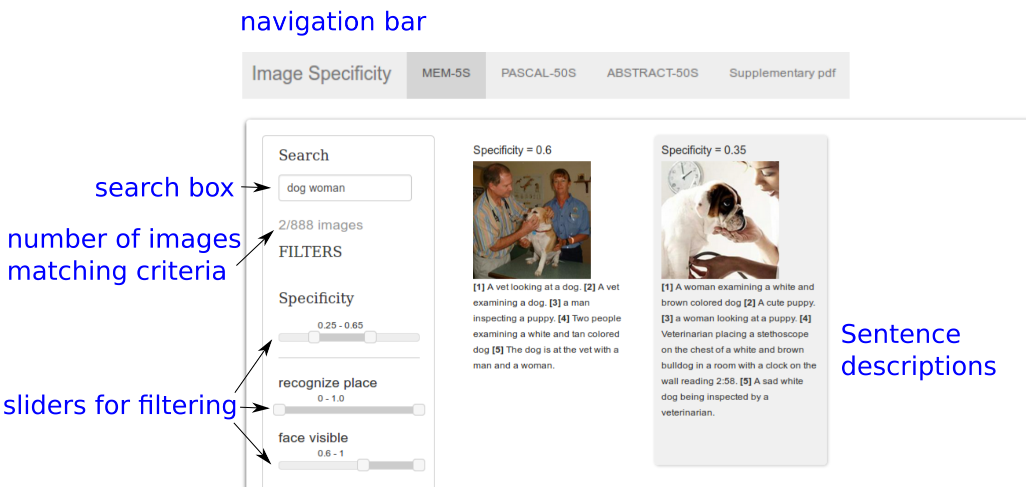


Figure 6. Dataset browser for exploring the datasets. Available on the authors' webpages.

## 4. Specificity vs. Memorability

In our paper, we have shown that specificity and memorability are correlated. However, they are distinct concepts and measure different properties of the image. In particular, we have shown that peaceful and picture-perfect scenes are negatively correlated with memorability but have no effect on specificity. In Figure 5, we show examples of images that are specific/not specific and memorable/not memorable. Note how outdoor scenes tend to be not very memorable but can have a reasonably high specificity score.

## 5. Website for exploring datasets

Here, we describe the website interface available on the authors' webpages that can be used to explore the datasets used in the paper. A navigation bar on top of the website allows users to switch between different datasets. Figure 6 shows how the search function can be used to look for sentences containing the words "dog" and "woman". Up to a maximum of 6 words can be added in the search box. Only whole words are matched. The reader should note that the website does not implement the text-based search algorithms discussed in the paper. It is meant for only browsing the datasets. Sliders on the left allow the user to filter images according to a range of scores that the images satisfy. All the criteria are combined using logical AND to display the filtered images. The number of images matching the search criteria gives the user an idea of how often two or more criteria are satisfied concurrently. The benefit of using such a website is that it can give the readers an intuition of the underlying data and factors that affect specificity. We have added sliders for the attributes that correlate most (top 10) and least (bottom 10) with specificity (for the MEM-5S dataset). It is also possible to filter by average length of the sentences and the memorability score.

## Glossary

**automated specificity** Specificity computed from image textual descriptions by averaging automatically computed sentence similarities (Section 3.1.2 in main paper) [1, 2] **human specificity** Specificity measured from image textual descriptions by averaging human-annotated sentence similarities (Section 3.1.1 in main paper) [1] **predicted specificity** Specificity computed from image features without any textual descriptions (Section 2) [1, 2]

## References

[1] S. Bird. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006. 2

[2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 1

[3] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011. 3

[4] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. 2

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011. 3

[6] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3016. IEEE, 2013. 1