

# Image Specificity (Supplementary material)

Mainak Jas  
Aalto University  
mainak.jas@aalto.fi

Devi Parikh  
Virginia Tech  
parikh@vt.edu

## 1. Scatter plots for correlations

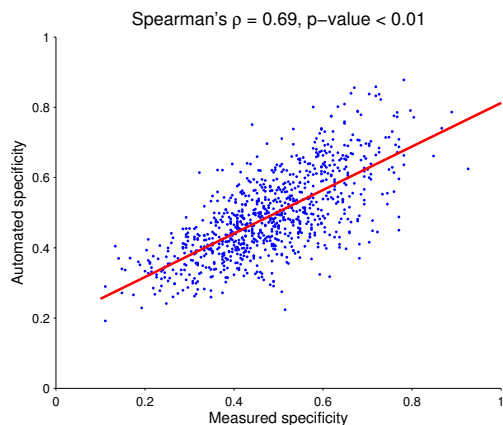


Figure 1. Correlation between **human-measured specificity** and **automated specificity** for the MEM-5S dataset.

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.

## 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

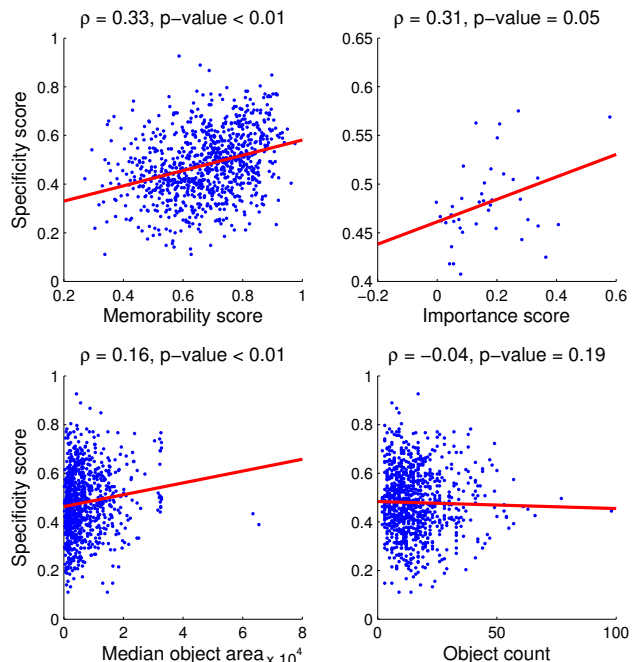


Figure 2. What makes an image specific? Memorable images, images with large objects and important object categories tend to be more specific. Number of annotated objects in an image does not correlate with specificity. Results are on the MEM-5S dataset.

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.* occurrence 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

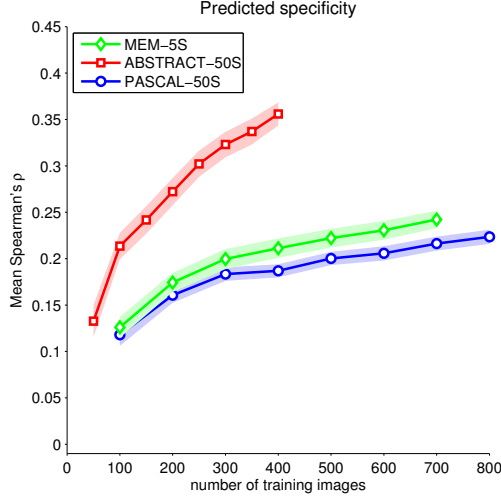


Figure 3. Spearman’s rank correlation between **predicted** and **automated specificity** for increasing number of training images (averaged across 50 random runs). **Automated specificity** (Section 3.1.2 in main paper) uses 5, 48 and 50 sentences per image for the three datasets, MEM-5S, ABSTRACT-50S 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.

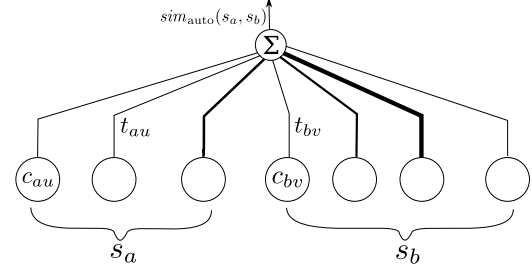
### 3. Detailed explanation of automated specificity computation

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

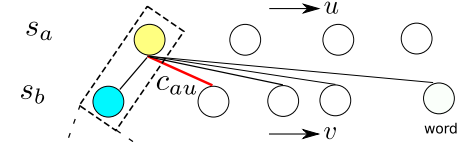
**A.** The **similarity between two sentences** is a weighted average of the contributions of each word with the TFIDF scores

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



**B.** The **contribution** is computed as the maximum similarity between a word and all words in the other sentence

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



**C.** **Similarity between two words** is the maximum similarity between all pairs of synsets they belong to

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

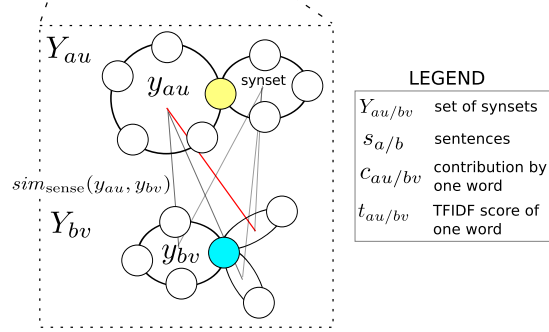


Figure 4. Illustration of our approach to compute automated sentence similarity.

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,

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

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  $s_b$ .

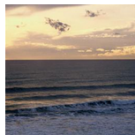
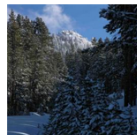



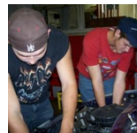




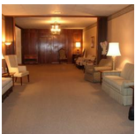

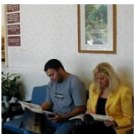
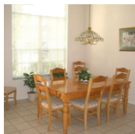








Specific, not memorable	<p>Specificity = 0.68 Memorability = 0.37</p>  <p>[1] A view of the ocean at sunset. [2] The sunset on the horizon of the ocean. [3] The waves are gently breaking on the shore. [4] The ocean is at mid-tide under the sunset. [5] Rolling waves under a sun set.</p>	<p>Specificity = 0.66 Memorability = 0.35</p>  <p>[1] A forest in the middle of a snowy mountainside. [2] There are many trees covered with snow. [3] A forest covered with fresh snowfall. [4] The mountain loomed in the distance over the snowy forest. [5] Snow covered sky blue scene with many trees</p>	<p>Specificity = 0.64 Memorability = 0.35</p>  <p>[1] some snow covered mountains. [2] This is a snowy mountain peak. [3] This is a snow-covered mountain. [4] A view of a snowy mountainside [5] A glacier between snow covered mountains.</p>	<p>Specificity = 0.63 Memorability = 0.39</p>  <p>[1] A mountain with a group of houses. [2] A village sits in the foothills of a rocky hillside. [3] a mountain behind some houses. [4] Many houses in front of a mountain. [5] A large mountain towering over a town.</p>			
	<p>Specificity = 0.89 Memorability = 0.66</p>  <p>[1] There is a lot of snow on the mountain. [2] There is a snow covered mountain. [3] A snow covered mountain. [4] A mountain with snow. [5] A snowy mountain.</p>	<p>Specificity = 0.87 Memorability = 0.69</p>  <p>[1] Two young men trying to fix their car together [2] Two men working on a car. [3] The men are working on the car. [4] Two men fixing a car [5] Two mechanics working on fixing a car.</p>	<p>Specificity = 0.85 Memorability = 0.9</p>  <p>[1] A person standing at a gun target range. [2] A man firing a pistol at a shooting range. [3] A man practicing how to shoot a gun with ear plugs on [4] The man is practicing on the shooting range. [5] A man shooting a gun at a shooting range.</p>	<p>Specificity = 0.8 Memorability = 0.86</p>  <p>[1] The backseat of a car. [2] The rear seat of a car. [3] A car seat in the back. [4] Leather seat in a car [5] the back seat of a car.</p>	<p>Specificity = 0.8 Memorability = 0.75</p>  <p>[1] A tall, twisting roller-coaster and blue skies [2] a rollercoaster. [3] A roller coaster [4] A view of a large roller coaster [5] A roller coaster.</p>	<p>Specificity = 0.78 Memorability = 0.81</p>  <p>[1] A bank vault sitting partially open. [2] The heavy door to a bank vault stands half open and an inner barred cage is visible within, along with some safety deposit boxes and a cushioned bench seat. [3] A bank vault with its door opened. [4] A large bank vault is standing open. [5] An open vault door.</p>	
	Specific, memorable	<p>Specificity = 0.4 Memorability = 0.64</p>  <p>[1] Beige upholstered furniture is placed close to the walls of a seating area that also seems to be a passageway in a hotel. [2] The inside of a lobby in a hotel. [3] A lobby area with lamps and furniture. [4] The long room is lined with neutral furniture. [5] A hallway with lamps, chairs and sofa well lit for the customers</p>	<p>Specificity = 0.4 Memorability = 0.8</p>  <p>[1] There is a covered bridge over water. [2] A bridge with water running underneath it. [3] Bridge over the water [4] a small shed. [5] A covered bridge over a river.</p>	<p>Specificity = 0.4 Memorability = 0.82</p>  <p>[1] A man and a woman reading magazines in a waiting room. [2] a man and a woman under a picture. [3] Two people reading on a couch. [4] Two people are sitting in chairs reading magazines. [5] A couple sitting in a waiting room.</p>	<p>Specificity = 0.4 Memorability = 0.8</p>  <p>[1] a table with a bouquet on it. [2] Wooden dining room set in a sunny room. [3] A dining room table. [4] A dining room table with matching chairs in a home. [5] A kitchen table under a light in front of a window.</p>	<p>Specificity = 0.4 Memorability = 0.64</p>  <p>[1] the inside of a building. [2] Hall with a glass ceiling [3] A giant hallway with chandeliers. [4] Things are hanging from the ceiling of this large room. [5] A large hall with many decorations hanging from the ceiling</p>	<p>Specificity = 0.4 Memorability = 0.82</p>  <p>[1] People are sitting down waiting. [2] A couple of kids in an airport. [3] some people in an airport. [4] People in a waiting room [5] A woman rests her head on another woman's shoulder in a waiting area.</p>
		<p>Specificity = 0.39 Memorability = 0.33</p>  <p>[1] A view of a multi-colored house [2] A house outside. [3] A white house with red trim and a brick chimney. [4] A house with a wall in front of it. [5] A two story house.</p>	<p>Specificity = 0.33 Memorability = 0.33</p>  <p>[1] A train station near some tracks. [2] An old building beside train tracks. [3] A building. [4] the outside of a pagoda. [5] A stone and timber railroad depot with a blue roof.</p>	<p>Specificity = 0.31 Memorability = 0.32</p>  <p>[1] A view of a large forest from a hill top. [2] The trees in the valley are changing colors. [3] A large open valley with fall colors. [4] The bushes spread over many miles of the desert. [5] Green, yellow mountain</p>	<p>Specificity = 0.24 Memorability = 0.34</p>  <p>[1] A deck with some people in front of a flight of stairs [2] A large area with weapons. [3] a battle ship with rockets. [4] People on a naval ship. [5] There are people looking at a cannon.</p>	<p>Specificity = 0.22 Memorability = 0.3</p>  <p>[1] A building in front of a mountain. [2] A modern house with windows facing the west. [3] A old rock pit and a building. [4] A building is near a beach with a log. [5] a building with some trees.</p>	<p>Specificity = 0.11 Memorability = 0.34</p>  <p>[1] Neon artwork suspended from the ceiling of an airport terminal. [2] A hocky rink. [3] Large empty room with shiny floors [4] the inside of a warehouse. [5] Two people are in a large area with televisions.</p>
		Not Specific, not memorable					

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 simi-

larity of that sentence with other sentences. So we have

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

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

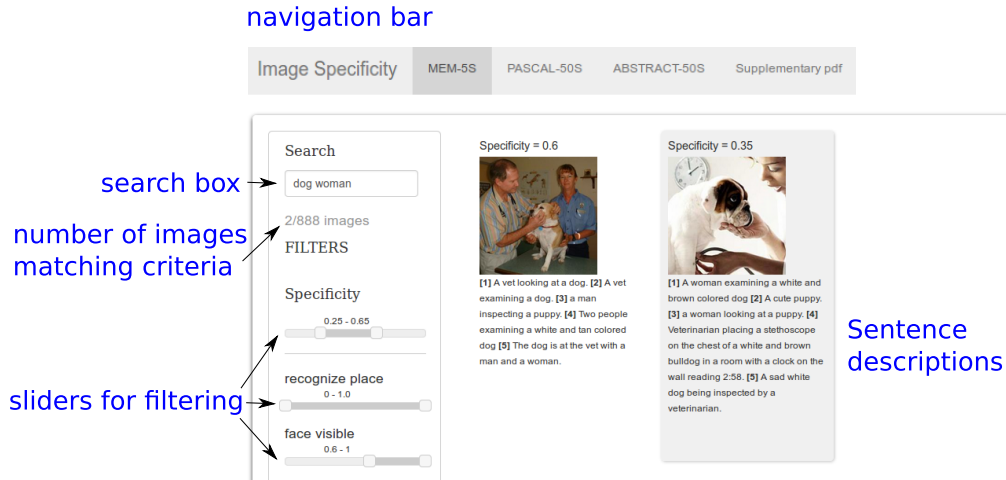


Figure 6. Dataset browser for exploring the datasets. Available on the authors’ webpages.

#### 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