Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval

Albert Gordo and Diane Larlus
Computer Vision group, Xerox Research Center Europe
firstname.name@xrce.xerox.com

Abstract

Querying with an example image is a simple and intuitive interface to retrieve information from a visual database. Most of the research in image retrieval has focused on the task of instance-level image retrieval, where the goal is to retrieve images that contain the same object instance as the query image. In this work we move beyond instance-level retrieval and consider the task of semantic image retrieval in complex scenes, where the goal is to retrieve images that share the same semantics as the query image. We show that, despite its subjective nature, the task of semantically ranking visual scenes is consistently implemented across a pool of human annotators. We also show that a similarity based on human-annotated region-level captions is highly correlated with the human ranking and constitutes a good computable surrogate. Following this observation, we learn a visual embedding of the images where the similarity in the visual space is correlated with their semantic similarity surrogate. We further extend our model to learn a joint embedding of visual and textual cues that allows one to query the database using a text modifier in addition to the query image, adapting the results to the modifier. Finally, our model can ground the ranking decisions by showing regions that contributed the most to the similarity between pairs of images, providing a visual explanation of the similarity.

1. Introduction

The task of image retrieval aims at, given a query image, retrieving all images relevant to that query within a potentially very large database of images. This topic has been heavily studied over the years. Initially tackled with bag-of-features representations, large vocabularies, and inverted files [61, 51], and then with feature encodings such as the Fisher vector or the VLAD descriptors [55, 31], the retrieval task has recently benefited from the success of deep learning representations such as convolutional neural networks that were shown to be both effective and computationally efficient for this task [64, 58, 25]. Among previous retrieval methods, many have focused on retrieving the exact same instance as in the query image, such as a particular landmark [56, 57, 32] or a particular object [51]. Another group of methods have concentrated on retrieving semantically-related images, where "semantically related" is understood as displaying the same object category [65, 8], or sharing a set of tags [23, 22]. This requires to make the strong assumption that all categories or tags are known in advance, which does not hold for complex scenes.

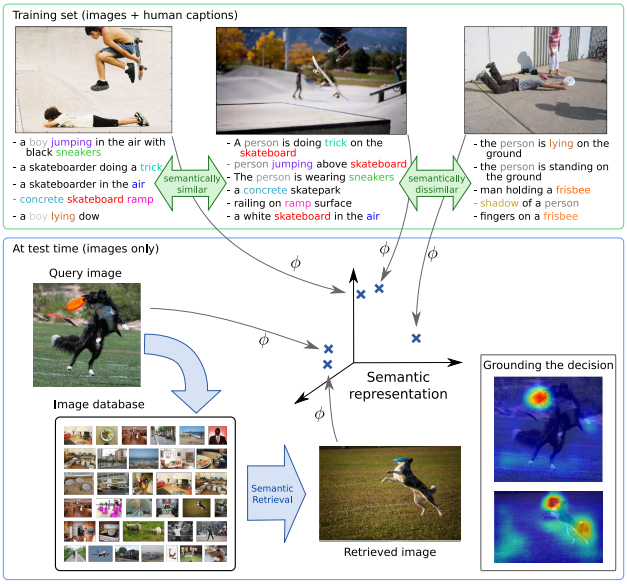

Figure 1. We tackle the **semantic retrieval** task. Leveraging the multiple human captions that are available for images of a training set, we train a semantic-aware representation that improves semantic visual search within a disjoint database of images that do not contain textual annotations. As a by-product, our method highlights regions that contributed the most to the decision.

In this paper we are interested in applying the task of *semantic retrieval* to query images that display realistic and complex scenes, where we cannot assume that all the object categories are known in advance, and where the inter-

action between objects can be very complex. Our first contribution is to *validate that the task of semantic retrieval is well-defined*, particularly in the presence of complex scenes (section 3). Although what different persons understand as a semantically similar scene is subject to interpretation, we show in a user study that there is a high level of consistency between different users.

Following the standard image retrieval paradigm that targets efficient retrieval within databases of potentially millions of images, we aim at learning a global and compact visual representation tailored to the semantic retrieval task that, instead of relying on a predefined list of categories or interactions, implicitly captures information about the scene objects and their interactions. However, directly acquiring enough semantic annotations from humans to train such a model may not be feasible. Our second contribution is to show that *a similarity function based on captions produced by human annotators, which we assume are available at training time, constitutes a good computable surrogate of the true semantic similarity*, and provides enough information to learn a semantic visual representation (section 4).

Our third contribution is a model that leverages the similarity between human-generated captions, *i.e*. privileged information available only at training time, to *learn how to embed images in a semantic space, where the similarity between embedded images is related to their semantic similarity* (section 5.1). Our experiments first show that learning a semantic representation significantly improves over a model pretrained on ImageNet. We also show that it can provide a visual explanation of the semantic similarity by highlighting regions that contributed the most to it.

Our last contribution (section 5.2) is an extension of the previous model that *leverages the image captions explicitly and learns a joint embedding for the visual and textual representations*. We show that this further improves the accuracy of the model, and, more importantly, this allows one to add text modifiers to the query in order to refine the query or to adapt the results towards additional concepts.

2. Related work

Image retrieval. Image retrieval has been mostly tackled as the problem of instance-level image retrieval [61, 51, 55, 31, 64, 58, 25], that focuses on the retrieval of the exact same instance as defined in standard benchmark datasets [56, 57, 32, 51]. Moving away from instances, some works have tackled visual search as the retrieval of images that share the same category label [7, 8] or a set of tags [29, 22]. These works still have a crude understanding of the semantics of a scene. On their synthetic dataset of abstract scenes, Zitnick and Parikh have shown that image retrieval can be greatly improved when detailed semantics is available [70]. Explicit modeling of a scene can be done with attributes [19, 17, 53, 41], object co-occurrences [47], or pairwise relationship between objects [12, 14, 43]. As the interaction between objects in a scene can be highly complex, going beyond simple pairwise relations, one extreme interface proposed by Johnson *et al*. [34] is to compare explicit scene graph representations instead of visual representations. One shortcoming of their method is that it requires the user to query with a full scene graph, which is a tedious process. We believe that querying with an image is a more intuitive interface.

A number of approaches have cast the task of image captioning as a retrieval problem, first retrieving similar images, and then transferring caption annotations from the retrieved images to the query image [28, 62, 18, 52]. Yet these methods use features that are not trained for the task, either simple global features [28], features pretrained on ImageNet [62] or complex features relying on object detectors, scene classifiers, etc. [18, 52]. We believe that the representation should be free of assumptions about the list of objects, attributes, and interactions one might encounter in the scene, and therefore, we learn these representations directly from the training data.

Joint embeddings of image and text. Many tasks require to jointly leverage images and natural text, such as zero-shot learning [4, 10], language generation [67, 35], multimedia retrieval [2, 3], image captioning [62, 16], and VQA [59, 45, 6]. A common solution is to build a joint embedding for textual and visual cues and to compare the modalities directly in that space. The first category of methods for joint embedding is based on CCA [26]. Recent methods using CCA include [22, 24, 39] and [5], a deep extensions of CCA. As an alternative to CCA, previous work has learned the joint embedding with a ranking loss. Among them, WSABIE [69] and DeViSE [20] learn a linear transformation of visual and textual features with a single-directional ranking loss. Some papers have proposed a bidirectional ranking loss [35, 36, 38, 62] possibly with additional constraints [68]. Deep methods have also been proposed for this task, based on deep Boltzman machines [63], auto-encoders [50], LSTMs [15], or RNNs [46]. These joint image and text embeddings are often used to do cross-modal queries, *i.e*. to retrieve image with textual queries and vice-versa [68].

In many of these works learning the joint embedding is, by itself, the end objective. This differs from our work, where the end task is to learn a visual embedding to retrieve images using a query image, and where the joint embedding is used to enrich the visual representation. From that point of view, a connection is also found with the privileged learning framework [66]: our improved representation is learned with privileged information in the form of semantic similarity measures provided by the captions that are present at training time. The work of Gomez *et al*. [21], in these same proceedings, follows a similar idea, leveraging text

from the Wikipedia to learn self-supervised visual embeddings aimed at classification, detection, and retrieval tasks.

3. User study

In this section we conduct a user study to acquire annotations related to the semantic similarity between images as perceived by users, and use those to show that the task of semantic retrieval in complex scenes is well-defined and that users tend to agree on their decisions. We also show that visual models pretrained on ImageNet, although better than random, do not reach a high agreement with the users, and that some form of training will be required to achieve good semantic retrieval results using only visual features.

Dataset. The computer vision community has made a recent effort in collecting and organizing large-scale datasets allowing for both training and benchmarking of cognitive scene understanding tasks: the MS-COCO dataset [42], the VQA dataset [6], that adds to MS-COCO a set of question/answer pairs related to the visual content of these images, and, more recently, the Visual Genome dataset [40], that is composed of 108k images with a wide range of annotations such as region level-captions, scene graphs, objects, and attributes. This dataset has been designed to evaluate tasks that go beyond image classification and that require to reason about the visual scenes. We adopt the Visual Genome dataset for our experiments, as it is well suited for the task of semantic visual search. We structure it into 80k images for training, 10k for validation, and 10k for test.

Methodology. Manually ranking a large set of images according to their semantic relevance to a query image is a very complex, tedious, and time-consuming task. Instead, to ease the task of the annotators, we consider the problem of triplet ranking: given a triplet of images, composed of one query image and two other images, we ask our users to choose the most semantically similar image to the query among the two options. To not bias the annotations towards any interpretation of semantic similarity, we keep the guidelines as open as possible, asking the users to choose, among the two displayed images, the one that "depicted the scene that was most similar to the scene in the query image". The users have the choice to choose one of the two images or to choose that both images were either equally relevant or not relevant to the query.

To construct the triplets we randomly sample query images and then choose two images that are visually similar to the query. This is achieved by extracting image features using ResNet-101 [27] pretrained on ImageNet (performing global average pooling after the last convolutional layer) and sampling two images from the 50 nearest neighbors to the query in the visual feature space. The motivation to choose visually similar images is that, in random image triplets, both images will most often be irrelevant to the query. Our study involves 35 annotators (13 women and 22 men), whose annotations spread over 3,000 image triplets. A common set of 50 triplets was answered by 25 users, and most triplets were annotated by at least two annotators. For every triplet we store three values: o_1 and o_2 encode the number of times the first (resp. second) image was chosen, and o_3 the number of times people did not pick any of the two images.

Inter-user agreement. We evaluate the agreement between users on this ranking task. We compute a score in a leave-one-user-out fashion, where the decisions of each user are compared against the decisions of all the other users. Given a user and a ranking question, the agreement score s is measured as the proportion of the remaining users that made the same choice as the user, weighted by the proportion of remaining users that made a decision on that triplet, *i.e.*, $s = w\frac{o_i - 1}{o_1+o_2-1}$, with $w = \frac{o_1+o_2-1}{o_1+o_2+o_3-1}$ and $i \in \{1, 2\}$ is the choice of the user. This score is only computed for triplets where both the user and at least one of the remaining users chose one of the images. The final agreement score for a particular user is the average of the per-triplet agreements. In average, the inter-user agreement score is 89.1, with a standard deviation of 4.6. This shows that people generally agree with each other on the semantic similarity ranking between two images. On the set of 50 images that was annotated by 25 users, we get a similar leave-one-out agreement score of 87.3 ± 4.5.

Agreement with visual representations. We now show that a model pretrained on ImageNet, with no further training, does not achieve a high agreement with the users. We consider an image representation based on the fully-convolutional ResNet-101 architecture [27]. Our representation follows the R-MAC [64, 25] architecture, where, after the convolutional layers from [27], one performs max-pooling over different grid regions of the image at different scales, normalizes the descriptors of each region independently using PCA with whitening, and finally aggregates and renormalizes the final output to obtain a descriptor of 2048 dimensions. These ResNet R-MAC descriptors can be compared using the dot product.

As in the inter-user agreement case, the agreement between a method and the users is measured as the proportion of users that agree with the ranking decisions produced by the method, weighted by the proportion of users that made a decision on that triplet, averaged through all the triplets with at least one human annotator. Under this setup, our visual baseline, the ResNet with R-MAC, obtains an agreement of 64.0, *cf*. Table 1. This agreement is higher than a random ranking of triplets (50.0 ± 0.8 over 5 runs), but significantly lower than the inter-user agreement, suggesting that training the visual models is necessary, and that, to that end, semantic annotations will be necessary.

Method	score
Human annotators	89.1 ± 4.6
Visual baseline: ResNet R-MAC	64.0
Object annotations	63.4
Human captions: METEOR	72.1
Human captions: word2vec + FV	70.1
Human captions: tf-idf	76.3
Generated captions: tf-idf	62.5
Random (x5)	50.0 ± 0.8

Table 1. Top row, inter-human annotation agreement on the image ranking task. Bottom rows: comparison between the semantic ranking provided by human annotators and several visual baselines and methods based on the Visual Genome annotations.

4. Proxy measures for semantic similarity

To learn a visual embedding that preserves the semantic similarity between images one would need a large number of annotated image triplets. Unfortunately, requiring human annotators to provide rankings for millions of triplets is not feasible. Instead, we propose to use a surrogate measure. Ideally, this surrogate measure should be efficient to compute and be highly correlated with the ranking given by the human annotators. To this end, we leverage the annotations of the Visual Genome dataset and study which measures yield a high correlation with the human annotators.

Our first representation leverages the objects contained in images. We consider the ground-truth object annotations provided with the Visual Genome dataset [40], that list all the objects present in one image and, when relevant, their WordNet [49] synset assignment. We build a histogram representation of each image, counting how many objects of each synset appear in that image, and weight the histograms with a tf-idf mechanism followed by ℓ_2 normalization. The final representations are compared with the dot product. As seen in Table 1, the agreement of this representation with the users is worse than the visual agreement. This shows that counting objects from a predefined list of categories and neglecting their interactions does not offer a good proxy for semantic similarity, and that more information is needed.

Motivated by this, we consider human captions as a proxy for semantic similarity. Our rationale is that the human annotators will have a bias towards annotating parts of the image that they deem important, and that these annotated parts will be the same that they use to decide if images are semantically similar or not. The Visual Genome dataset contains, on average, 50 region-level captions per image annotated by different users, and this redundancy should further help to capture subtle semantic nuances. Consequently we leverage the provided region-level captions to build several textual representations of the images.

An intuitive way to compare image captions is to use METEOR [13], a similarity between text sentences typically used in machine translation that has also been used as a standard evaluation measure for image captioning [11]. To compare two sets of region-level captions X and Y from two images, we perform many-to-many matching with a (non-Mercer [44]) match kernel of the form

$$K(X,Y) = \frac{1}{|X|+|Y|}\Big(\sum_{x\in X}\max_{y\in Y} M(x,y) + \sum_{y\in Y}\max_{x\in X} M(x,y)\Big).$$

Note that this requires to evaluate up to thousands of pairs of sentences to compare two images, which may take up to a few seconds for images with more than a hundred captions. Therefore, the scalability of this approach is limited.

To avoid the scalability problem, one option is to merge all the words of all the captions of an image into a single set of words. This sacrifices the structure of the sentences but allows to use other methods based on bags of words. We experiment with two of them. The first one follows [30] and computes a Fisher vector [54] (FV) of the word2vec [48] representations of the captions' words. The semantic similarity between two captioned images is the dot product between the two ℓ_2-normalized FV representations. The second one is a *tf-idf* weighting of a bag-of-words (BoW) followed by ℓ_2 normalization, that can also be compared using the dot product. Contrary to the METEOR metric, these two last approaches produce not only a similarity but also a vectorial representation of the text that can potentially be used during training. All learning involved in these representations (vocabulary of 46881 words, idf weights, Gaussian mixture model for the word2vec-based Fisher vector, etc.) is done on our training partition of the Visual Genome dataset.

We compute the agreement score of all these methods by comparing their decision to the users', and report results in Table 1. We observe that the region-level captions provided by human annotators are very good predictors of the semantic similarity between two images, much better than the visual baseline ones. Of these, the tf-idf BoW representation is best, outperforming METEOR and word2vec on this task. Consequently, this is the representation we leverage to train a better visual representation in the next section. As a comparison, we also experimented with automatically-generated captions [1, 67] instead of user-generated captions. The score of the automatic captions is significantly lower, highlighting the importance of using human captions for training.

5. Learning visual representations

In the previous section we have shown that human generated captions capture the semantic similarity between images. Here we propose to learn a global image representation that preserves this semantic similarity (Section 5.1). We then extend our method to explicitly embed the visual and textual representations jointly (Section 5.2).

5.1. Visual embedding

Our underlying visual representation is the ResNet-101 R-MAC network discussed in Section 3. This network is designed for retrieval [64] and can be trained in an end-to-end manner [25]. Our objective is to learn the optimal weights of the model (the convolutional layers and the projections in the R-MAC pipeline) that preserve the semantic similarity. As a proxy of the true semantic similarity we leverage the tf-idf-based BoW representation over the image captions. Given two images with captions we define their proxy similarity as the dot product between their tf-idf representations.

To train our network we propose to minimize the empirical loss of the visual samples over the training data. If q denotes a query image, d^+ a semantically similar image to q, and d^- a semantically dissimilar image, we define the empirical loss as as $L = \sum_q \sum_{d^+,d^-} L_v(q, d^+, d^-)$, where

$$L_v(q, d^+, d^-) = \frac{1}{2} \max(0, m - \phi_q^T \phi_+ + \phi_q^T \phi_-), \quad (1)$$

m is the margin and $\phi : \mathcal{I} \rightarrow \mathbb{R}^D$ is the function that embeds the image into a vectorial space, *i.e.* the output of our model. We slightly abuse the notation and denote $\phi(q)$, $\phi(d^+)$, and $\phi(d^-)$, as ϕ_q, ϕ_+, and ϕ_-. We optimize this loss with a three-stream network as in [25] with stochastic optimization using ADAM [37].

To select the semantically similar d^+ and dissimilar d^- images we evaluated two approaches. In the first one we directly sample them such as that $s(q, d^+) > s(q, d^-)$, where s is the semantic similarity between two images, computed as the dot product between their tf-idf representations, as above. However, we observed this sampling strategy not to improve the visual representation. We believe this is because this strategy optimizes the whole ranking at once, and in particular tries to produce a correct ranking for images that are all very relevant, and for images that are all irrelevant, simply based on visual information. This is an extremely challenging task that our model was not able to correctly learn. Instead, for the second approach, we adopt a hard separation strategy. Similar to other retrieval works that evaluate retrieval without strict labels (*e.g.* [33]), we consider the k nearest neighbors of each query according to the similarity s as relevant, and the remaining images as irrelevant. This significantly simplifies the problem, as now the goal is to separate relevant images from irrelevant ones given a query, instead of producing a global ranking. Despite the hard thresholding, we observe this approach to learn a much better representation. Note that this thresholding is done only at training time, not at testing time. In our experiments we use $k = 32$, although other values of k led to very similar results. To reduce the impact of this thresholding the loss could also be scaled by a weight involving the semantic similarity, similar to the WARP loss [69], although we did not explore this option in this work. Finally, note that the human captions are only needed at training time to select image triplets, and are not used at test time.

5.2. A joint visual and textual embedding

In the previous formulation, we only used the textual information (*i.e.* the human captions) as a proxy for the semantic similarity in order to build the triplets of images (query, relevant and irrelevant) used in the loss function. In this section, we propose to leverage the text information in an explicit manner during the training process. This is done by building a joint embedding space for both the visual representation and the textual representation. For this we define two new losses that operate over the text representations associated with the images:

$$L_{t1}(q, d^+, d^-) = \frac{1}{2} \max(0, m - \phi_q^T \theta_+ + \phi_q^T \theta_-), \quad (2)$$

$$L_{t2}(q, d^+, d^-) = \frac{1}{2} \max(0, m - \theta_q^T \phi_+ + \theta_q^T \phi_-). \quad (3)$$

As before, m is the margin, $\phi : \mathcal{I} \rightarrow \mathbb{R}^D$ is the visual embedding of the image, and $\theta : \mathcal{T} \rightarrow \mathbb{R}^D$ is the function that embeds the text associated with the image into a vectorial space of the same dimensionality as the visual features. We define the textual embedding as $\theta(t) = \frac{W^T t}{\|W^T t\|_2}$, where t is the ℓ_2-normalized tf-idf vector and W is a learned matrix that projects t into a space associated with the visual representation.

The goal of these two textual losses is to explicitly guide the visual representation towards the textual one, which we know is more informative. In particular, the loss in Eq. (2) enforces that text representations can be retrieved using the visual representation as a query, implicitly improving the visual representation, while the loss in Eq. (3) ensures that image representations can be retrieved using the textual representation, which is particularly useful if text information is available at query time. All three losses (the visual and the two textual ones) can be learned simultaneously using a siamese network with six streams – three visual streams and three textual streams. Interestingly, by removing the visual loss (Eq. (1)) and keeping only the joint losses (particularly Eq. (2)), one recovers a formulation similar to popular joint embedding methods such as WSABIE [69] or DeViSE [20]. In our case, however, retaining the visual loss is crucial as we target a query-by-image retrieval task, and removing the visual loss leads to inferior results. We also note that our visual loss shares some similarities with the structure-preserving loss of [68], although they tackle the very different task of cross-modality search (*i.e.* sentence-to-image and image-to-sentence retrieval).

6. Experiments

This section validates the representations produced by our proposed semantic embeddings on the semantic re-

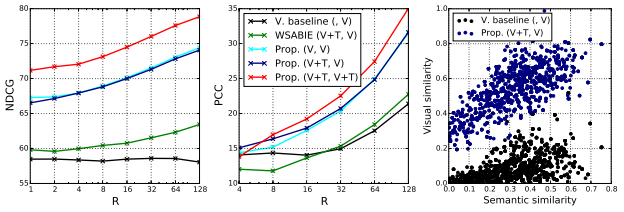

Figure 2. Left and center: NDCG and PCC achieved by the different models as a function of the number of retrieved images R, where the ground truth is determined by the tf-idf similarity. Right: correlation between the ground truth tf-idf similarity and the visual similarity of the baseline and trained models.

trieval task. We quantitatively evaluate them in two different scenarios. In the first one, we evaluate how well the learned embeddings are able to reproduce the semantic similarity surrogate based on the human captions. In the second scenario, we evaluate our models using the triplet-ranking annotations acquired from users (Section 3), by comparing how well our visual embeddings agree with the human decisions on all these triplets. Then, we propose an experiment that shows which parts of the images led to the matching score. Finally we illustrate how, by leveraging the joint embedding, the results retrieved for a query image can be altered or refined using a text modifier.

6.1. Experimental details

Implementation details. Our visual model is based on the ResNet-101 architecture [27] (pretrained on ImageNet) for the convolutional layers followed by the R-MAC pooling, projection, aggregation, and normalization pipeline [64]. We resize all our images preserving the aspect ratio such as that the largest side is of 576 pixels, and use two scales for the R-MAC pooling. To extract textual features we encode the captions using tf-idf. We stem the words using the Snowball stemmer from NLTK [9].

Our models are learned with a batch size of 64 triplets (sextuples depending on the setup) using the ADAM optimizer with an initial learning rate of 10^{-5}, which is reduced to 10^{-6} after 8k iterations. To mine triplets for training, we follow a similar approach to [25, 58]. We first randomly sample $N = 500$ images. For each of those N samples, we sample 9 relevant images according to the ground truth. This produces a pool of 5000 images, where at least 500 of them have at least 9 relevant images in the pool. Then we extract their features using the current state of the model, and prepare all possible triplets of query image, relevant image, and irrelevant image involving the images in the pool, and where the query is only sampled from the first N images. Finally, the 100 triplets with the largest loss for every query and positive pair are selected as potential candidates to be sampled and used for updating the model. This mining process is repeated after $t = 64$ updates of the model.

	US	NDCG AUC	PCC AUC
Text oracle			
Caption Tf-idf	76.3	100	100
Query by image			
Random (x5)	50.0 ± 0.8	10.2 ± 0.1	-0.2 ± 0.7
Visual baseline (, V)	64.0	58.4	16.1
WSABIE (V+T, V)	67.8	61.0	15.7
Proposed (V, V)	76.9	70.1	20.7
Proposed (V+T, V)	77.2	68.8	21.1
Query by image + text			
Proposed (V+T, V+T)	78.6	74.4	22.5

Table 2. Comparison of the proposed methods and baselines evaluated according to User-study (US) agreement score, AUC of the NDCG and PCC curves (*i.e.* NDCG AUC and PCC AUC).

Metrics. We benchmark our proposed models with two metrics that evaluate how well they correlate with the tf-idf proxy measure, which is the task we optimize for, as well as with the user agreement metric proposed in Section 3. Although the latter corresponds to the exact task that we want to address, the metrics based on the tf-idf similarity provide additional insights about the learning process and allow one to crossvalidate the model parameters. We evaluate our approach using normalized discounted cumulative gain (NDCG) and Pearson's correlation coefficient (PCC). Both measures are designed to evaluate ranking tasks. PCC measures the correlation between ground-truth and predicted ranking scores, while NDCG can be seen as a weighted mean average precision, where every item has a different relevance – in our case, the relevance of one item with respect to the query is the dot product between their tf-idf representations. To evaluate our method in the validation or test splits we choose 1k images from the split, that are used as queries, and use them to rank all the 10k images in the split. The query image is removed from the results. Finally, since we are particularly interested in the top results, we do not report results using the full list of 10k retrieved images. Instead, we report NDCG and PCC after retrieving the top R results, for different values of R, and plot the results.

Methods and baselines. We evaluate different versions of our embedding. We denote our methods with a tuple of the form ($\{$V, V+T$\}$, $\{$V, V+T$\}$). The first element denotes whether the model was trained using only visual embeddings (V), *cf*. Eq. (1), or joint visual and textual embeddings (V+T), *cf*. Eq. (1)-(3). The second element denotes whether, at test time, one queries only with an image, using its visual embedding (V), or with an image and text, using its joint visual and textual embedding (V+T). In all cases, the database consists only of images represented with visual embeddings, with no textual information.

Our approach is compared to our ResNet-101 R-MAC baseline, pretrained on ImageNet, with no further training, and to a WSABIE-like model, that seeks a joint embedding

optimizing the loss in Eq. (2), but does not explicitly optimize the visual retrieval goal of Eq. (1).

6.2. Results and discussion

We start by discussing the effect of training in the task of simulating the semantic similarity surrogate function. Figure 2 presents the results using the NDCG@R and PCC@R metrics for different values of R.

Our first observation is that all forms of training improve over the ResNet baseline. Of these, WSABIE is the one that obtains the smallest improvement, as it does not optimize directly the retrieval end goal and only focuses on the joint embedding. All methods that optimize the end goal obtain significantly better accuracies. The second observation is that, when the query consists only of one image, training our model explicitly leveraging the text embeddings – models denoted with (V+T, V) – brings quantitative improvement over (V,V) only on some of the metrics. However, this joint training allows one to query the dataset using both visual and textual information – (V+T, V+T). Using the text to complement the visual information of the query leads to significant improvements.

In Table 2 we evaluate these methods on the human agreement score. For context, we also report the area under the curve (AUC) of the NDCG and PCC curves. As with NDCG and PCC, learning the embeddings brings substantial improvements in the user agreement score. In fact, all of our trained models actually outperform the score of the tf-idf over human captions, which was used as a "teacher" to train our model, following the learning with privileged information terminology. Our model leverages both the visual features as well as the tf-idf similarity during training, and, as such, it is able to exploit the complementary information that they offer. Using text during testing improves agreement with users, and brings considerable improvements in the NDCG and PCC metrics. Additionally, having a joint embedding can be of use even if quantitative results do not improve, for instance for refining the query, see Figure 5.

Grounding the decisions. We leverage recent visualization techniques to highlight the regions of a pair of images that contribute the most to their similarity. We follow Grad-CAM [60], that displays the aggregated activations of the last convolutional layer weighted with the gradient of the loss for a target class. In our case, instead of using the gradients with respect to a specific class, we use the gradients with respect to the top $k = 5$ dimensions of the final signatures that contributed the most to their similarity. Figure 3 displays pairs of images, where the key regions that most contributed to the similarity are highlighted. Please note how the same image can highlight different regions depending on with which image it has been matched with.

Qualitative retrieval results. Figure 4 compares the visual baseline with our trained method (V+T,V), where our method retrieves more semantically meaningful results, such as horses on the beach or newlyweds cutting a wedding cake. Figure 5 shows the effect of text modifiers. The embedding of the query image is combined to the embeddings of textual terms (that can be added or subtracted to the representation) to form a new query with an altered meaning that is able to retrieve different images, and that is only possible thanks to the joint embedding of images and text.

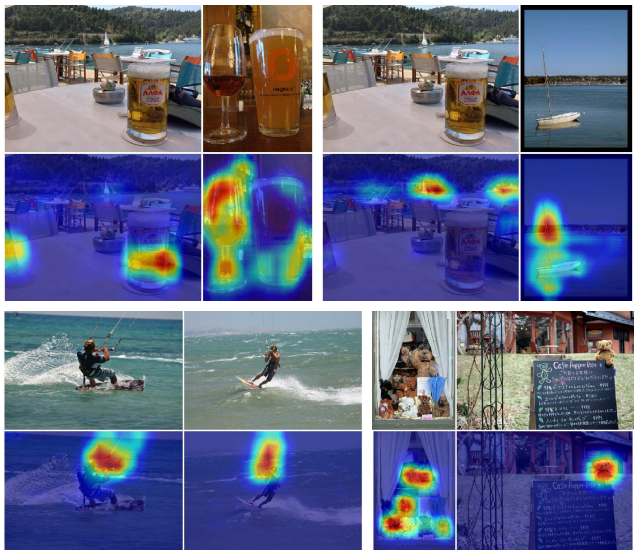

Figure 3. Grounding the decisions. For every pair of images we highlight the parts that contributed the most to their learned visual similarity. Different parts of the same image are highlighted depending on the image it is matched to.

7. Conclusions

In this work we focus on the task of semantic image retrieval, where, given a query image, the goal is to retrieve images that depicts similar scenes. To this end we conducted a user study and showed that i) users typically agree on the task of semantically ranking images, and ii) these ranks can be accurately predicted by exploiting human-annotated captions. We leveraged these annotations to learn a visual embedding of the images and showed that this visual embedding predicts very well the human ranking preferences, even better than the human caption proxy we trained with. Our models can also provide visual explanations about why a pair of images is similar. Finally, our joint visual and textual model can leverage text modifiers to refine the meaning of a query image, providing exciting new ways to query image databases.

Acknowledgments. We would like to thank Florent Perronnin for fruitful discussions and all of our 35 annotators.

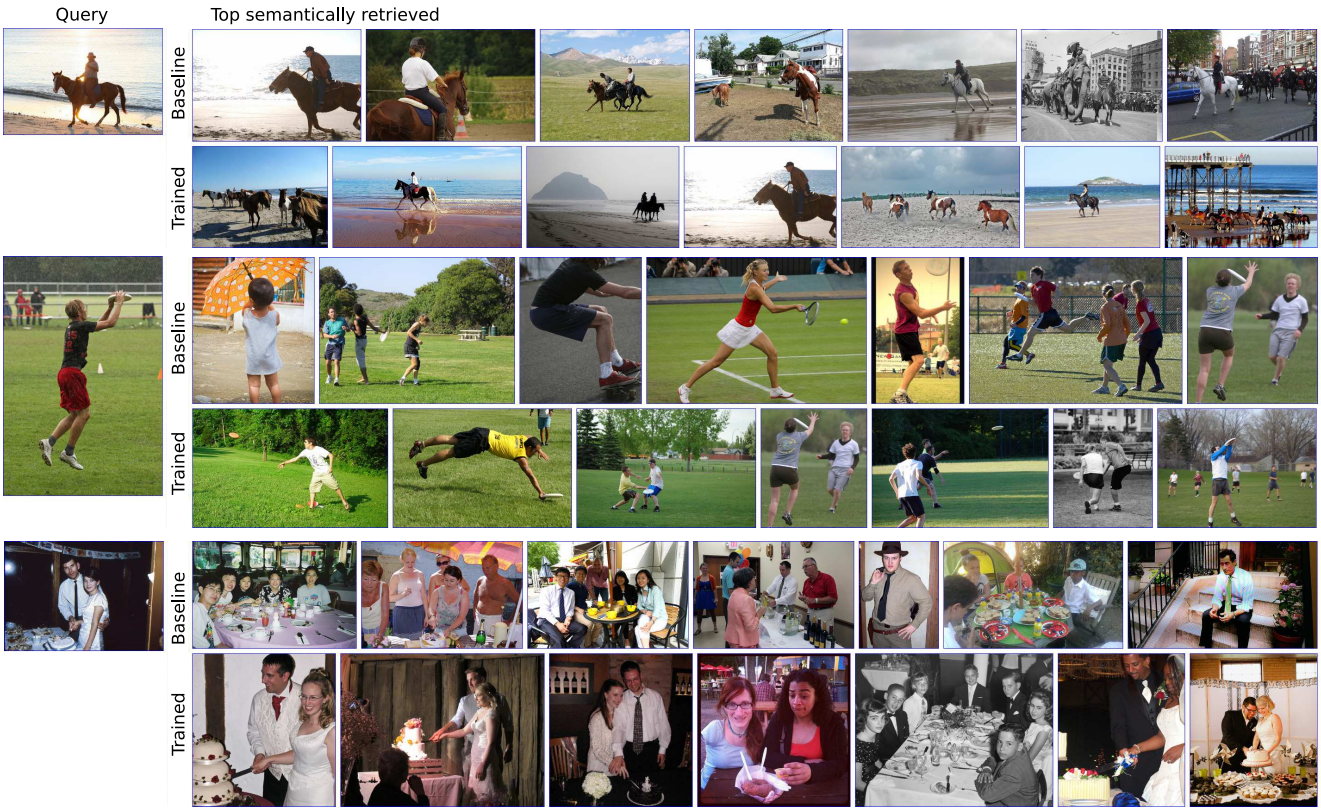

Figure 4. Qualitative results. For every block of images, left: query image. top: top-7 images with the representation pretrained on ImageNet, bottom: top-7 images with our learned representation with the (V+T,V) model.

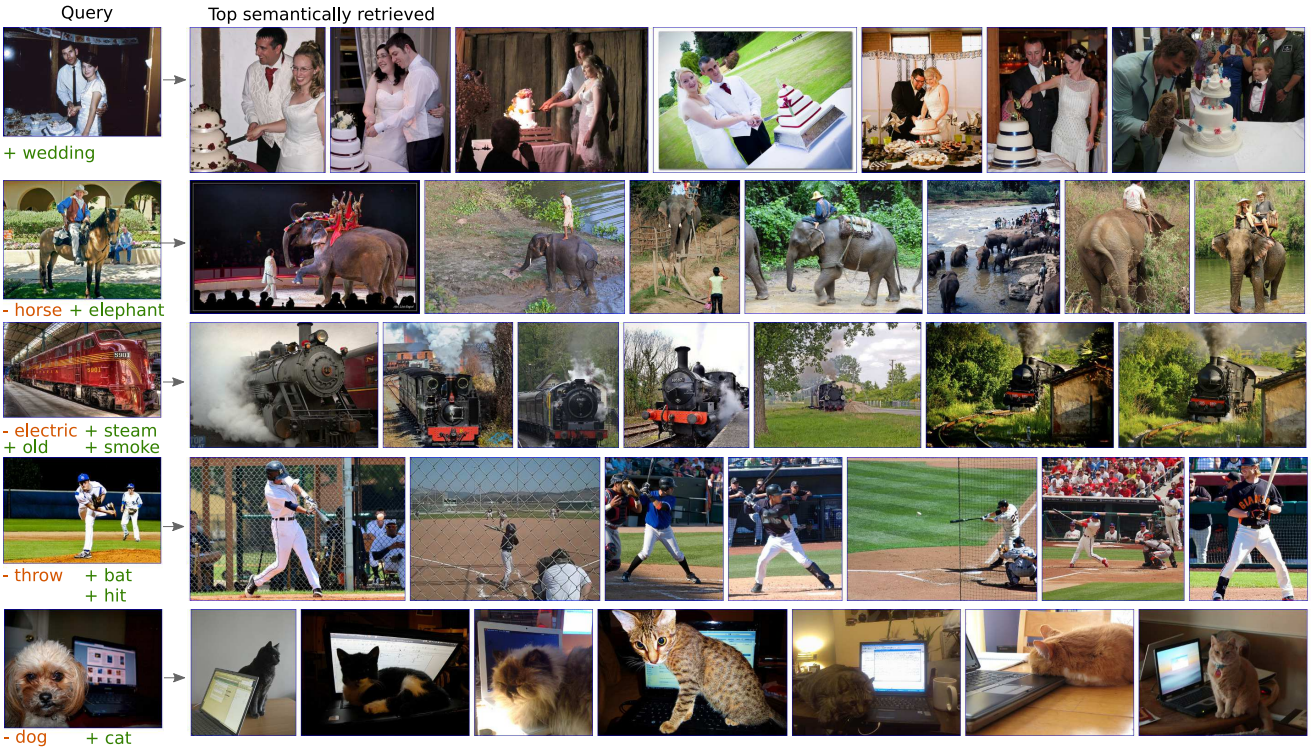

Figure 5. For a set of query images, we use a text modifier as additional query information (concepts are added or removed) to bias the results. Note that the first query is the last one from Figure 4 refined with additional text.

References

[1] http://t-satoshi.blogspot.fr/2015/12/image-caption-generation-by-cnn-and-lstm.html. 4

[2] J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J. Renders. Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications*, 42(1):31–56, 2009. 2

[3] J. Ah-Pine, G. Csurka, and S. Clinchant. Unsupervised visual and textual information fusion in cbmir using graph-based methods. *ACM Transactions on Information Systems*, 33(2):9:1–9:31, 2015. 2

[4] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele. Multi-cue zero-shot learning with strong supervision. In *CVPR*, 2016. 2

[5] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, 2013. 2

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2, 3

[7] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2

[8] A. Bergamo, L. Torresani, and A. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*. 2011. 1, 2

[9] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. 2009. 6

[10] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, 2016. 2

[11] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 4

[12] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2

[13] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *EACL Workshop on Statistical Machine Translation*, 2014. 4

[14] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 2011. 2

[15] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[16] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015. 2

[17] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2

[18] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010. 2

[19] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007. 2

[20] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *NIPS*, 2013. 2, 5

[21] L. Gomez, Y. Patel, M. Rusiñol, D. Karatzas, and C. V. Jawahar. Self-supervised learning of visual features thorugh embedding images into text topic spaces. In *CVPR*, 2017. 2

[22] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014. 1, 2

[23] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *TPAMI*, 35(12):2916–2929, 2013. 1

[24] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014. 2

[25] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016. 1, 2, 3, 5, 6

[26] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. 2

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6

[28] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47(1):853–899, May 2013. 2

[29] S. J. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 2012. 2

[30] M. Jain, J. van Gemert, T. Mensink, and C. Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, 2015. 4

[31] H. Jégou and O. Chum. Negative evidences and cooccurences in image retrieval: The benefit of pca and whitening. In *ECCV*. 2012. 1, 2

[32] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*. 2008. 1, 2

[33] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *TPAMI*, 2011. 5

[34] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[35] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2014. 2

[36] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 2

[37] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[38] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. 2

[39] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015. 2

[40] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016. 3, 4

[41] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2014. 2

[42] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 3

[43] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 2

[44] S. Lyu. Mercer kernels for object recognition with local features. In *CVPR*, 2005. 4

[45] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2

[46] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). In *ICLR*, 2015. 2

[47] T. Mensink, E. Gavves, and C. G. M. Snoek. COSTA: Co-Occurrence Statistics for Zero-Shot Classification. In *CVPR*, 2014. 2

[48] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 4

[49] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995. 4

[50] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011. 2

[51] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006. 1, 2

[52] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2

[53] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 2

[54] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 4

[55] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010. 1, 2

[56] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 1, 2

[57] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008. 1, 2

[58] F. Radenovic, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016. 1, 2, 6

[59] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 2

[60] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 7

[61] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1, 2

[62] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218, 2014. 2

[63] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *JMLR*, 15:2949–2980, 2014. 2

[64] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*, 2016. 1, 2, 3, 5, 6

[65] L. Torresani, M. Szummer, and A. Fitzgibbon. Learning query-dependent prefilters for scalable image retrieval. In *CVPR*, 2009. 1

[66] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. In *IJCNN*, 2009. 2

[67] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2, 4

[68] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 2, 5

[69] J. Weston, S. Bengio, and N. Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011. 2, 5

[70] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 2