



# A Structured Listwise Approach to Learning to Rank for Image Tagging

Jorge Sánchez<sup>1,2(✉)</sup>, Franco Luque<sup>1,2</sup>, and Leandro Lichtensztein<sup>3</sup>

<sup>1</sup> CONICET, Córdoba, Argentina

{jorge.sanchez, franco.luque}@unc.edu.ar

<sup>2</sup> Universidad Nacional de Córdoba, Córdoba, Argentina

<sup>3</sup> Deep Vision AI Inc., Córdoba, Argentina

leandro.lich@deepvisionai.com

**Abstract.** With the growing quantity and diversity of publicly available image data, computer vision plays a crucial role in understanding and organizing visual information today. Image tagging models are very often used to make this data accessible and useful. Generating image labels and ranking them by their relevance to the visual content is still an open problem. In this work, we use a bilinear compatibility function inspired from zero-shot learning that allows us to rank tags according to their relevance to the image content. We propose a novel listwise structured loss formulation to learn it from data. We leverage captioned image data and propose different “tags from captions” schemes meant to capture user attention and intra-user agreement in a simple and effective manner. We evaluate our method on the COCO-Captions, PASCAL-sentences and MIRFlickr-25k datasets showing promising results.

**Keywords:** Learning to rank · Zero-shot learning · Image tagging  
Visual-semantic compatibility · Multimodal embedding

## 1 Introduction

In the past decade, we have witnessed a tremendous growth in the quantity and diversity of media resources, especially images and videos. With all this information being stored and shared across social and media platforms, the ability to search and to organize such data efficiently is a problem of great practical importance.

One of the main characteristics of social media data is its multimodal nature: images and videos are frequently associated with user generated textual descriptions (brief captions, tags, hashtags, etc.), providing complementary information not necessarily present or apparent in the visual domain. Besides enriching and complementing the visual information, the textual information can be used to index the data, facilitating their access and analysis.

For computer agents trying to organize the data in an autonomous manner, being able to automatically generate and rank tags and labels is an important task. An alternative is to rely on a large pool of image classifiers and/or

object detectors. However, using off-the-shelf classifiers for image tagging leads to results that does not resemble tags that a human user would choose to describe the content of an image. For instance, predictions cast by models trained on ImageNet [1] correspond to leaf nodes in the WordNet [2] lexical ontology and tend to be overly specific. Moreover, image datasets annotated with object categories usually consist of a restricted set of labels that reflects the presence or absence of objects in images, disregarding other objects or visual properties that sometimes are more relevant from the perceptual point of view. For instance, if we look at some of the examples in Table 1 we see that annotations based on a fixed set of object categories often miss important visual information (e.g. the mirror in the first image) or give the same “relevance” to all objects, irrespective of their role in explaining the semantic content of the scene (e.g. apple vs. refrigerator in the third image).

Using annotations from tagging datasets, e.g. NUS-WIDE [3] or MIRFlickr [4], where annotations correspond to actual tags generated by users on the Flickr website, has also some difficulties. First, tags might be unrelated to the actual visual content of the images, e.g. pictures tagged with the camera brand/model they were captured with. Second, the set of possible tags, although richer, is still restricted to a closed set of possible words.

In this work, we aim at learning a visual-semantic compatibility function that allow us to rank textual descriptions (tags) according to their relevance to the image content, without restricting ourselves to a fixed vocabulary at test time. We propose a novel structured listwise ranking loss that encodes explicitly the relevance of the tags to the actual content on the visual domain. Our approach assumes the availability of a training set of image and ranked tags pairs. We build such a training set by leveraging captioned image data like the COCO Captions dataset [5]. Using multiple caption annotations, we propose a simple and yet effective method for the extraction of image tags and to rank them according to their relevance on explaining the visual scene. Our approach is based on the following observations: *(i)* common words chosen by different users are good candidates for image tags (intra-user agreement), and *(ii)* terms named earlier in a sentence are visually more relevant than those named at the end (user attention). We run extensive experimental evaluations on three different datasets.

The rest of the paper is organized as follows. First, we present an overview of related work. Then, we describe our model and propose different methods to infer ranked tags from image captions. Last, experiments, results and discussion are presented to conclude the paper.

## 2 Related Work

We now review related work on research areas that we believe are closely related to our work, namely: image tagging, zero-shot learning (ZSL) and embeddings for multi-modal data.

*Tag Assignment and Refinement.* We focus on tagging methods that use information provided by images and tags. A common intuition in such methods is that visually similar images should share a similar set of tags. In the TagProp model [6], tags for a test image are predicted based on a weighted sum of the annotations of the most visually similar images on the training set. In [7], a distance metric learning scheme is proposed, exploiting both image and tag information in a transductive way. The learning formulation involves a triplet-based max-margin objective, solved by stochastic gradient descent (SGD). TagCooccur+ [8] combines visual and tag information into a relevance score based on the co-occurrence frequencies both on the visual and textual domain. RobustPCA [9] factorizes the image-tag association matrix using a low rank decomposition with an  $\ell_1$  sparsity constraint. [10] proposes an approach based on Markov random walks on a graph built from image similarities and image-tag associations. The method proved to scale to very large datasets. [11] uses both images and tags to build a graph. Learning takes place on the structure of this graph based on samples and a pseudo-relevance measure. Different from other graph-based approaches, edge weights are also learned, allowing to minimize the effect of uninformative tags and visual words. We refer the reader to [12] for an extensive review of the problems and different approaches proposed in the literature.

*Zero-Shot Learning and Recognition.* Zero-shot learning aims at recognizing object categories that might not have been seen during training [13–16]. In the literature, there exists two formulations of this problem. In the original formulation, it is assumed that train and test classes are disjoint. In the generalized version of the zero-shot learning problem (GZSL) this assumption is relaxed and the sets are allowed to overlap. This problem has shown to be more difficult than ZSL. We refer the reader to [17] and [18] for recent surveys on the topic.

Most approaches dealing with either ZSL or GZSL assume that images and class labels can be encoded as points in some vector spaces, e.g. feature vectors extracted based on a pre-trained network [19, 20] and word/attribute embeddings derived from side information [13, 21–23]. A compatibility function between (the representations of) images and class labels is then learned from training data. In this case, the most common approach is based on the use of bilinear forms [16] and which we also follow in this work.

*Multi-modal Embeddings.* In the fields of computer vision (CV) and natural language processing (NLP), the use of convolutional neural networks (CNN) and distributional semantic models has led to major advances. The combination of both modalities [24–26] has shown great potential on several linguistic tasks. For instance, [26] extends the skip-gram model of [22, 23] by taking into account the visual information associated with a restricted set of words. The model showed good performance on a variety of semantic benchmarks. [27] uses autoencoders to learn grounded meaning representations from images and textual data. Experimental results on word similarity and word categorization showed that multimodal information improves over unimodal counterparts. Even though our

approach does not learn a multimodal representation explicitly, we do constraint the visual and textual spaces to be aligned in terms of their semantics.

### 3 Our Model

Given an image  $x \in \mathcal{X}$  and a tag  $y \in \mathcal{Y}$ , our goal is to learn a function  $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  that provides us with a score regarding the compatibility between the semantics of  $y$  and the visual content in  $x$ . Following [16], we model  $s$  as a simple bilinear map:

$$s_{(\psi, \phi)}(x, y; W) = \psi(x)^T W \phi(y), \quad (1)$$

where  $\psi : \mathcal{X} \rightarrow \mathbb{R}^D$  and  $\phi : \mathcal{Y} \rightarrow \mathbb{R}^E$  denote image and word embeddings, respectively. For a given choice of  $(\psi, \phi)$ , our goal is to learn  $W \in \mathbb{R}^{D \times E}$  from a suitable set of training samples.

In our work, we assume the availability of a training set  $\mathcal{D} = \{(x_n, Y_n)\}_{n=1}^N$  consisting of images  $x_n$  and ordered tag-sets  $Y_n = \{y_1^n, \dots, y_{|Y_n|}^n\}$ , with  $r(y_1^n; x_n) \geq \dots \geq r(y_{|Y_n|}^n; x_n)$  for a given relevance measure  $r$ , i.e.  $(Y_n, \geq_{r(\cdot; x_n)})$  is a partially ordered set. In what follows, we assume  $r$  is given, and thus the preference order of the tags in  $Y_n$  is known, for all  $n$ . Later, in Sect. 3.2 we will discuss different formulations for  $r$  based on some consistencies observed in human-generated captions, when different annotators are asked to describe the content of an image.

#### 3.1 Learning Formulation

We consider loss functions of the form:

$$L(W; r) = \sum_{n=1}^N \ell(\hat{Y}(x_n), Y_n), \quad (2)$$

with  $Y_n = \{y_1^n, \dots, y_{|Y_n|}^n\}$  the tags for  $x_n$  as ranked by  $\geq_{r(\cdot; x_n)}$  and  $\hat{Y}(x_n)$  the same set of tags but ranked according to  $\geq_{s_{(\psi, \phi)}(\cdot, x_n; W)}$ , i.e. the loss function  $\ell$  encodes the cost of predicting an order for the tags in  $Y_n$  different from that induced by the ground-truth measure  $r$ .

**Structured Joint Embedding (SJE).** First, we follow [16] and consider loss functions of the form:

$$\ell_{SJE}(x, Y) = \max_{1 \leq i \leq |Y|} [\Delta(1, i) + s_{(\psi, \phi)}(x, y_i) - s_{(\psi, \phi)}(x, y_1)]_+, \quad (3)$$

with  $[z]_+ \equiv \max(0, z)$ . We explore two different formulations for the structured term  $\Delta : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$  over relative orders, namely:

$$\Delta^{(I)}(k, k') = 1 - \delta_{k, k'} \quad (4)$$

$$\Delta^{(II)}(k, k') = 1 - (k' - k + 1)^{-1}, \quad k' \geq k \quad (5)$$

Equation (4) corresponds to the structured loss of [16] when we assume the top-ranked tag to be the only relevant tag for the image under consideration. Equation (5) considers the relative order of the tags. Its effect on Eq. (3) is to ensure that the ground truth tag (top-1 in  $Y$ ) receives a higher score than the other tags in the list.

**Listwise Structured Joint Embedding (ListSJE).** One of the shortcomings of the previous formulation is that it penalizes wrong associations only w.r.t to the most relevant tag in the list. Inspired by the likelihood loss formulation of Xia *et al.* [28], we propose the following cost function:

$$\ell_{ListSJE}^{K_{top}}(x, Y) = \sum_{k=1}^{K_{top}} \sum_{k < i \leq |Y|} [\Delta(k, i) + s_{(\psi, \phi)}(x, y_i) - s_{(\psi, \phi)}(x, y_k)]_+ \quad (6)$$

where we assume  $K_{top} \leq |Y|$  and  $\Delta(k, k') \equiv \Delta^{(II)}(k, k')$ . Compared to Eq. (3), the loss given by Eq. (6) does also considers the relative ordering of the tags within the set of relevant annotations.

### 3.2 From Image Captions to Ranked Tags

Our formulation assumes the availability of a dataset  $\mathcal{D} = \{(x_n, Y_n)\}_{n=1}^N$  of images, each of which is annotated with a list of tags sorted by decreasing relevance according to a given measure  $r$ . Note that, given a set of possible tags for an image, different definitions of  $r$  will lead to different preference relations for the tags in the set. Defining  $r$  is also challenging, since the notion of “visual relevance” for a tag might be influenced by external factors which are rather subjective and difficult to grasp, e.g. user intentions, pre-existing knowledge, social context, etc. Our approach is to build  $\mathcal{D}$  from data by leveraging existing datasets like COCO [5] where, besides object categories, each image is annotated with a set of 5 different captions describing its visual content. Some example annotations are shown in Table 1. Next, we describe different approaches to extract a list of ranked tags from the captions available for each image.




Let  $\mathcal{C}(x) = \{c_1, \dots, c_Q\}$  be the set of captions corresponding to image  $x$ . We denote as  $t(c_i) \equiv t_i = \{w : w \in c_i \text{ and } w \text{ is a noun}\}$  the set of nouns extracted from  $c_i \in \mathcal{C}(x)$ . Also, let  $\text{loc}(w; c)$  denote the relative location of word  $w$  within  $c$ , e.g.  $\text{loc}(w = \text{“dog”}; c = \text{“The dog bites.”}) = 2/3$  as “dog” is the second word on a three-word sentence, and  $\text{count}(w; c)$  the number of times word  $w$  appears in  $c$ . We define the following scores:

$$r_{loc}(w) = \max_{c \in \mathcal{C}(x)} \{1 - \text{loc}(w; c) : w \in c\} \quad (7)$$

$$r_{freq}(w) = \frac{\text{count}(w; c_1) + \dots + \text{count}(w; c_Q)}{|t_1| + \dots + |t_Q|} \quad (8)$$

The first is a proxy for visual attention (under the hypothesis that objects that are mentioned earlier in a sentence are those which are more relevant to the

**Table 1.** COCO example annotations (best viewed with magnification) and ranked tags obtained by the relevance score of Eq. (9) on nouns and  $\alpha = 0.5$ .

|   | Categories  | Captions  | Tag   | $r_{0.5}^{nn}$   |
|---|---|---|---|--|
|  | clock   | <ol style="list-style-type: none"> <li>1. A brown mirror hanging on the wall.</li> <li>2. there is a mirror where u can see the reflection of a clock</li> <li>3. a clock in the reflection of a mirror</li> <li>4. A mirror on a wall reflecting a wooden clock.</li> <li>5. A mirror with a reflection of a clock in it.</li> </ol>   | mirror<br>clock<br>reflection<br>wall   | 0.633<br>0.580<br>0.425<br>0.371   |
|  | backpack<br>dog<br>person<br>surfboard  | <ol style="list-style-type: none"> <li>1. Two people walking with surf boards and two dogs.</li> <li>2. People with surf boards walking from the shoreline accompanied by dogs on a sunny day.</li> <li>3. A group of people and dogs carry their surfboards in hand.</li> <li>4. Two boys with dogs carry surfboards down the beach.</li> <li>5. Two people standing next to a river holding surfboards.</li> </ol>                              | people<br>surf<br>group<br>boy<br>board<br>dog<br>surfboard<br>shoreline<br>river<br>beach<br>hand<br>day   | 0.591<br>0.483<br>0.481<br>0.473<br>0.452<br>0.441<br>0.318<br>0.304<br>0.223<br>0.123<br>0.106<br>0.085 |
|  | apple<br>bottle<br>bowl<br>orange<br>oven<br>refrigerator<br>sink<br>wine glass | <ol style="list-style-type: none"> <li>1. A bright red retro refrigerator in a mostly white kitchen</li> <li>2. Cooking utensils and bowls hanging on a rack above an oven sitting in a kitchen with a refrigerator, sink and a window.</li> <li>3. a kitchen with a fridge and a sink below a window</li> <li>4. a kitchen with a sink a refrigerator and a window</li> <li>5. This kitchen has a red refrigerator and a black stove.</li> </ol> | kitchen<br>cooking<br>utensil<br>refrigerator<br>rack<br>retro<br>sink<br>fridge<br>oven<br>window<br>stove | 0.568<br>0.523<br>0.502<br>0.391<br>0.377<br>0.373<br>0.368<br>0.341<br>0.314<br>0.118<br>0.114          |

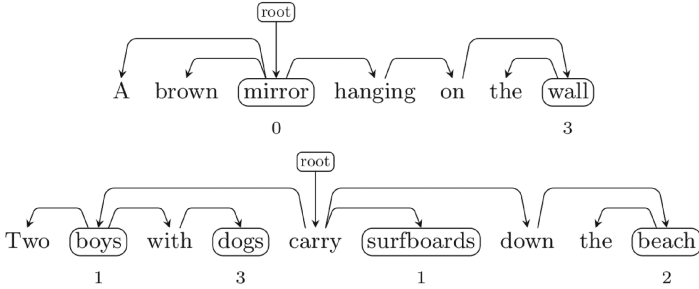
visual scene) while the second acts as a proxy for the agreement between different users and the terms they choose when asked to describe the content of a scene. We consider a simple combination of these scores and define a parameterized measure as follows:

$$r_{\alpha}(w) = \alpha r_{freq}(w) + (1 - \alpha) r_{loc}(w), \quad 0 \leq \alpha \leq 1 \tag{9}$$

Here, the hyper-parameter  $\alpha$  controls the strength of the frequency score w.r.t the location score.

We also consider two variations of the above based on the parse tree of each caption, namely:

1. Use compound nouns instead of only nouns, e.g. {“soccer field”} instead of {“soccer”, “field”}.
2. Use a syntactic version of the relative location score: the relative distance of the word to the root in the syntactic dependency tree of the caption. Relevant nouns do not necessarily appear first in a sentence, but they always appear at the top of dependency trees. In particular, main nouns are at the root of nominal phrases, the most common caption type, and subject/object nouns



**Fig. 1.** Syntactic dependency trees for two COCO captions. Edges point from parents to children. Nouns are highlighted and annotated with their (absolute) distances to the root word. The first caption is a nominal phrase with root noun “mirror”. The second caption is a declarative sentence with nouns “boys” and “surfboards” as subject and object respectively. Here, “dogs” occurs before “surfboards” but is less relevant according to the tree.

are directly attached to the root verb of declarative sentences. Examples of these are shown in Fig. 1. As a drawback, a natural language parser must be used, so errors made by the parser propagate to the score.

We denote the different combinations as  $r_{\alpha}^{nn}$ ,  $r_{\alpha}^{cn}$ ,  $r_{\alpha}^{nn-syntactic}$  and  $r_{\alpha}^{cn-syntactic}$ , where the superscripts *nn* and *cn* denote nouns and compound-nouns, respectively.

We note that, choosing a particular form of relevance completely defines the training set  $\mathcal{D}$  on which future models will be trained on. In what follows, when we refer to “a model” we refer not only to the loss function used to train the  $W$  matrix in Eq. (1) but also to the dataset (set of image and ranked tags pairs) used to train it.

## 4 Experiments

In this section we present experimental results regarding the models described above. We first present the datasets used in our experiments and the experimental setup we followed. Next, we discuss the benefits and limitations of the proposed approach and how it compares to other approaches in the literature.

### 4.1 Datasets

In our experiments, we use three different datasets: COCO Captions [5], PASCAL-sentences [29] and MIRFlickr-25k [4] that we describe next.

**COCO.** We use the training and validation sets of the 2014 release of the dataset, consisting of 123k images ( $\sim 83$ k for training and  $\sim 40$ k for validation) annotated with 5 different human descriptions each. In all our experiments, the

bilinear model of Eq. (1) is trained on the train set of the COCO dataset. To tune the parameters (learning rate, number of epochs, etc.) we use a subset of the training data. We use the validation set to evaluate different aspects of the model.

**PASCAL-Sentences.** The dataset contains 1k images from the PASCAL VOC 2008 Challenge [30], each of which was annotated with 5 different captions. We use the PASCAL-sentences dataset as an independent set on which to test our model. Note that, although the images from COCO and PASCAL are similar in the sense that both datasets focus on generic object recognition in natural scenes (no iconic views), the number of objects per image differs considerably. For instance, around 10% for COCO images contain a single object while this number increases to 60% for the images in PASCAL VOC [5].

**MIRFlickr-25k.** The dataset contains 25k images collected from Flickr together with the tags that users assigned to them. There are 1386 different tags, with an average of 8.94 tags per image. Besides providing generic user tags, the images were also manually annotated for a set of 24 different concepts. A second round of annotations was performed for 14 of the original concepts, where the images were deemed relevant for a given concept only if a significant part of the concept appeared in the image.

## 4.2 Experimental Setup

**Textual Features.** In our experiments, we use pretrained word2vec<sup>1</sup> [22,23], GloVe<sup>2</sup> [31] and fastText<sup>3</sup> [32] word embeddings. Nouns are lemmatized before computing the vectors. For compound nouns, we compute the average vector of the lemmatized constituent words. We use spaCy<sup>4</sup> [33] to process caption sentences and extract nouns, compound nouns, their lemmatized versions and syntactic dependency trees.

**Visual Features.** We use VGG [19] and ResNet [20] convolutional architectures to extract visual features from the images. Additionally, we also implemented a simplified version of the multiscale R-MAC feature extractor [34] from the retrieval literature, where we do not include the PCA whitening step after the region-level feature pooling operation. Visual features are extracted from the penultimate fully connected layer of a pre-trained VGG-19 or ResNet-152 architecture. For R-MAC we consider three different scales (1, 2 and 4) and pool features from activations of the last convolutional layer of a VGG-16 network. We L2-normalize the max pooled features, average them and re-normalize the resulting vector.

<sup>1</sup> <https://code.google.com/archive/p/word2vec/>.

<sup>2</sup> [https://spacy.io/models/en#en\\_core\\_web\\_md](https://spacy.io/models/en#en_core_web_md).

<sup>3</sup> <https://fasttext.cc/docs/en/english-vectors>.

<sup>4</sup> <https://spacy.io/>.



**Model Training.** To train our model we use mini-batch gradient descent with a batch size of 16 over 10 epochs. We use the Adam optimizer with an initial learning rate of 0.0001. All models were implemented in PyTorch v0.4.0 [35]. To train our models we used 3 NVIDIA GTX 1080Ti cards on an Intel Xeon machine @ 2.6 GHz with 64 GB of RAM. Training a single bilinear model took approximately 2 h.

### 4.3 Experimental Results

In this section, we evaluate and discuss different aspects of our model. These evaluations are carried out on the COCO and PASCAL-Sentences datasets. Next, we compare the performance of our approach with other methods proposed in the literature on the challenging MIRFlickr-25k dataset.

**Tags from Captions.** In Sect. 3.2 we presented a simple approach to extract a set of tags from the captions describing a given visual scene. These tags correspond to the nouns (or compound nouns) extracted from each image caption. As such, they can all be considered as “relevant” to the actual visual content. In our approach, however, what matters is not the relevance score of any particular tag but the order in which they appear in the annotation list. Instead of trying to predict a ground truth order, we rank the different approaches according to their ability to lead to predictable tags. That is, we train a bilinear model as in Eq. (1) for each of the losses in Sect. 3.1. We chose the best tag generation scheme based on the performance on the validation set of the COCO dataset. For these experiments, we rely on VGG-19 and word2vec as image and tag features, respectively. We use precision@1 and precision@5 as performance metrics. Tables 2, 3 and 4 show results for the different formulations and parameter  $\alpha$ , for each of the losses presented in Sect. 3.1. Best results are highlighted in bold. For  $\ell_{ListSJE}$  we set  $K_{top} = 5$ .

From the tables we see that, when considering p@1, the models based on simple nouns and  $\alpha = 0.75$ ,  $r_{0.75}^{nn}$ , perform best for all loss functions. Among them, the  $\ell_{SJE}$  formulation with the  $\Delta^{(II)}$  structured loss leads to the best p@1 score on the COCO validation set. This behavior changes when we consider p@5 as the evaluation metric. In this case, the model based on compound nouns leads to the best performance for all losses. We also note that, for p@5, setting  $\alpha = 0$  gives the best results. Interestingly, in these models the word frequencies play no role. Only relative location scores are considered.

Also interesting is that the best performance is observed for the listwise formulation of Eq. (6).<sup>5</sup> For the SJE loss of Eq. (3),  $\Delta^{(II)}$  is preferred over  $\Delta^{(I)}$ . In what follows, we focus on the  $r_{0.75}^{nn}$  and  $r_{0.0}^{cn}$  relevance measures.

<sup>5</sup> In preliminary experiments, we explored the use of the tag scores, as given by the relevance measure  $r$ , directly into the structured loss term in Eqs. (3) and (6). However, we did not observe any improvement w.r.t to the simpler formulations given by Eqs. (4) and (5).

**Table 2.** Tag predictability measured on the COCO validation set, using VGG-19 and word2vec features and the loss function of Eq. (3) with  $\Delta = \Delta^{(I)}$ .

|     | Method                     | $\alpha$      |        |        |               |        |
|-----|----------------------------|---------------|--------|--------|---------------|--------|
|     |                            | 0.0           | 0.25   | 0.5    | 0.75          | 1.0    |
| p@1 | Nouns                      | 0.4621        | 0.5752 | 0.6332 | <b>0.6457</b> | 0.5744 |
|     | Nouns (syntactic)          | 0.4108        | 0.6203 | 0.6257 | 0.6350        | 0.5730 |
|     | Compound-nouns             | 0.4435        | 0.5561 | 0.6095 | <b>0.6155</b> | 0.5018 |
|     | Compound-nouns (syntactic) | 0.4540        | 0.6019 | 0.6086 | 0.6140        | 0.5035 |
| p@5 | Nouns                      | <b>0.6876</b> | 0.6814 | 0.6788 | 0.6786        | 0.6634 |
|     | Nouns (syntactic)          | 0.6812        | 0.6791 | 0.6808 | 0.6782        | 0.6628 |
|     | Compound-nouns             | 0.6952        | 0.6915 | 0.6920 | 0.6868        | 0.6438 |
|     | Compound-nouns (syntactic) | <b>0.7015</b> | 0.6977 | 0.6973 | 0.6880        | 0.6438 |

**Table 3.** Tag predictability measured on the COCO validation set, using VGG-19 and word2vec features and the loss function of Eq. (3) with  $\Delta = \Delta^{(II)}$ .

|     | Method                     | $\alpha$      |        |        |               |        |
|-----|----------------------------|---------------|--------|--------|---------------|--------|
|     |                            | 0.0           | 0.25   | 0.5    | 0.75          | 1.0    |
| p@1 | Nouns                      | 0.4810        | 0.5848 | 0.6395 | <b>0.6527</b> | 0.5865 |
|     | Nouns (syntactic)          | 0.4391        | 0.6257 | 0.6309 | 0.6389        | 0.5859 |
|     | Compound-nouns             | 0.4638        | 0.5664 | 0.6161 | <b>0.6243</b> | 0.5205 |
|     | Compound-nouns (syntactic) | 0.4755        | 0.6069 | 0.6137 | 0.6195        | 0.5191 |
| p@5 | Nouns                      | <b>0.7162</b> | 0.7118 | 0.7031 | 0.7003        | 0.6747 |
|     | Nouns (syntactic)          | 0.7064        | 0.7024 | 0.7029 | 0.6981        | 0.6752 |
|     | Compound-nouns             | <b>0.7230</b> | 0.7169 | 0.7124 | 0.7055        | 0.6537 |
|     | Compound-nouns (syntactic) | 0.7200        | 0.7119 | 0.7109 | 0.7037        | 0.6537 |

**Influence of the Parameter  $K_{top}$ .** Next, we focus on the listwise ranking loss of Eq. (6). In particular, we evaluate the influence of the parameter  $K_{top}$  and different word embeddings on tag predictability. Results are shown in Fig. 2 for  $r_{0.75}^n$  and  $r_{0.0}^n$ . Considering the choice of word embedding, fastText and word2vec exhibit a similar performance, outperforming GloVe vectors for all values of  $K_{top}$ . When comparing different image features, VGG-19 shows better performance than ResNet-152 and R-MAC features for both  $r_{0.75}^n$  and  $r_{0.0}^n$  measures for all values of the parameter. In what follows, we choose VGG-19 and word2vec features as the best configuration.

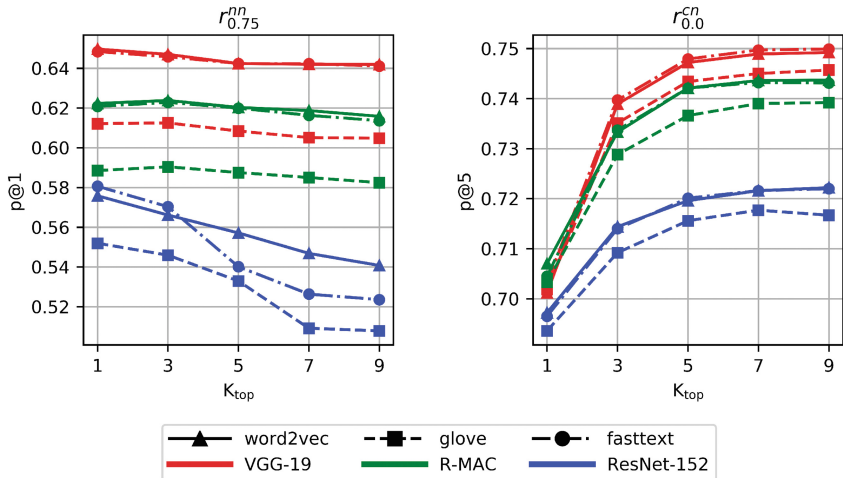
**Results on PASCAL-Sentences.** We now turn to the evaluation of our model on the PASCAL-sentences dataset. Figure 3 shows precision@ $k$  for different values of  $k$  for  $(\psi, \phi) = (\text{VGG-19}, \text{word2vec})$  and the ranking loss of Eq. (6). We also consider a varying number of *distractors*, i.e. tags sampled at random from

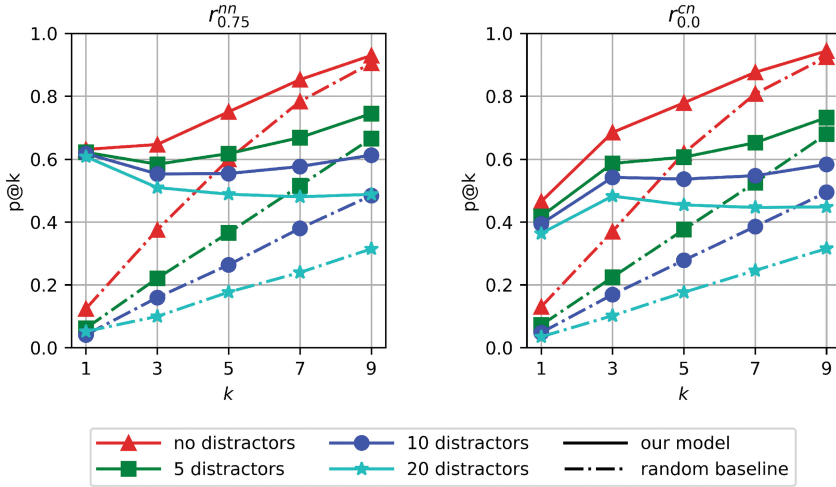
**Table 4.** Tag predictability measured on the COCO validation set, using VGG-19 and wor2vec features and the loss function of Eq. (6) with  $K_{\text{top}} = 5$ .

| Method |                            | $\alpha$      |        |               |               |        |
|--------|----------------------------|---------------|--------|---------------|---------------|--------|
|        |                            | 0.0           | 0.25   | 0.5           | 0.75          | 1.0    |
| p@1    | Nouns                      | 0.4660        | 0.5619 | 0.6250        | <b>0.6424</b> | 0.5791 |
|        | Nouns (syntactic)          | 0.4284        | 0.6085 | 0.6166        | 0.6274        | 0.5797 |
|        | Compound-nouns             | 0.4488        | 0.5391 | 0.5972        | <b>0.6123</b> | 0.5092 |
|        | Compound-nouns (syntactic) | 0.4722        | 0.5897 | 0.5979        | 0.6027        | 0.5098 |
| p@5    | Nouns                      | 0.7415        | 0.7411 | <b>0.7417</b> | 0.7385        | 0.6899 |
|        | Nouns (syntactic)          | 0.7302        | 0.7322 | 0.7329        | 0.7322        | 0.6897 |
|        | Compound-nouns             | <b>0.7472</b> | 0.7468 | 0.7461        | 0.7405        | 0.6646 |
|        | Compound-nouns (syntactic) | 0.7392        | 0.7391 | 0.7382        | 0.7335        | 0.6639 |

the set of all tags extracted from all the images in the dataset which are different from those of the query. Distractors are only sampled at test time. Additionally, we show the performance of a random ranker (dashed lines).

From the figure we observe that precision increases with  $k$ . This is to be expected since we rank a fixed pool of potential tags per image. However, the relative gain w.r.t to the random ranker is greater for a larger number of distractors. For instance, for the model trained on the  $r_{0.75}^{nn}$  setting and  $k = 5$  the gain in performance is +25%, +68%, +110% and +176% for 0, 5, 10 and 20 distractors per image, respectively. It is also interesting to note that for  $k = 1$ ,

**Fig. 2.** Precision@{1,5} for  $r_{0.75}^{nn}$  (left) and  $r_{0.0}^{cn}$  (right) as a function of  $K_{\text{top}}$  in Eq. (6), measured on the validation set of the COCO dataset (best viewed in color). (Color figure online)



**Fig. 3.** Precision@k on the PASCAL-sentences dataset using VGG-19 and word2vec features (best viewed in color). (Color figure online)

there is a large gap between the noun and compound-noun based systems, with the former giving +0.2 absolute improvement w.r.t the latter. For larger values of  $k$ , the difference decreases significantly.

**Results on MIRFlickr-25k.** In this section we compare the performance of our approach against other methods proposed in the literature on the MIRFlickr-25k dataset. We report mean average precision (MAP) and image-centered mean average precision (MiAP) on the set of 14 more restrictive (“relevant” set) tags as in [12]. We compare our method to TagProp and the SVM-based method proposed by the authors in [6]. We also compare to some of the methods reported in [12], namely: the UserTags baseline and two of the CNN-based models (TagCooccur+ [8] and RobustPCA [9]) trained on a 100k images dataset. These methods were shown to be among the top performing on the task of tag assignment and refinement [12].

For the experiments on this dataset, we build an image-tag affinity matrix whose entries are the average of two normalized similarity terms:

$$s_1(x, y) = \frac{\phi(x)^T W \psi(y)}{\|\phi(x)\| \|W\psi(y)\|}, \quad s_2(x, y) = \frac{\phi(x)^T \phi_y}{\|\phi(x)\| \|\phi_y\|} \quad (10)$$

where  $\phi_y$  denotes the average feature vector for the images in the training set tagged with term  $y$ . In the above,  $s_1$  is a normalized (on the image embedding space) version of the similarity score of Eq. (1) while  $s_2$  encodes the intuition similar images should share similar tags. To compute  $\phi_y$ , we generated train and test splits as in [6] and divided the dataset by taking every second image for training and the rest for testing. The matrix  $W$  was trained as before using the

ranked tags derived from the train set of the COCO dataset. Results are shown in Table 5.

Compared to the handcrafted feature-based methods of [6] (SVM and TagProp) our approach compares favorably. Although this is to be expected (we rely on more robust visual features), we opted to present the original results of [6] for reference. Note, however, that both TagProp and SVM achieve a better MAP score than CNN+TagCooccur+ [8] which is based on VGG-16 convolutional features. Compared to CNN+RobustPCA [9], our models are behind in terms of MAP but compare favorably in terms of MiAP.

**Table 5.** Results on the MIRFlickr-25k set.

| Method   | MAP   | MiAP  |
|--|-------|-------|
| TagProp (Rank) [6]                             | 0.404 | -     |
| SVM [6]  | 0.466 | -     |
| UserTags [12]                                  | 0.263 | 0.204 |
| CNN+TagCooccur+ [8, 12]                        | 0.381 | 0.277 |
| CNN+RobustPCA [9, 12]                          | 0.627 | 0.376 |
| Our ( $r_{0.75}^{nn}$ , $K_{\text{top}} = 1$ ) | 0.521 | 0.389 |
| Our ( $r_{0.0}^{cn}$ , $K_{\text{top}} = 7$ )  | 0.514 | 0.370 |

## 5 Conclusions

In this paper, we proposed a new method to learn a visual-semantic compatibility based on a structured listwise ranking loss formulation. Since there is no dataset containing per-image ranked tags in the literature – which is required to train our model – we take advantage of images captions from publicly available datasets and proposed several methods to automatically extract a list of tags from image captions, sorted according to its relevance to the visual content of the scene. Based on the this, we were able to train models that compare favorably to some other methods proposed in the literature, showing promising results. In future work, we want to explore different tag inference mechanisms as well as to include explicit models of visual attention, integrating the visual and semantic feature generation into an end-to-end architecture.

**Acknowledgments.** This work was supported in part by grants PICT 2014-1651 and 2016-0118 from ANPCyT, Argentinean Ministry of Education, Culture, Science and Technology. This work used Nabucodonosor Cluster from CCAD-UNC, which is part of SNCAD, Argentina.

## References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
2. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
3. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval, p. 48. ACM (2009)
4. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 39–43. ACM (2008)
5. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48)
6. Verbeek, J., Guillaumin, M., Mensink, T., Schmid, C.: Image annotation with TagProp on the MIRFlickr set. In: Proceedings of the International Conference on Multimedia Information Retrieval, pp. 537–546. ACM (2010)
7. Wu, P., Hoi, S.C.H., Zhao, P., He, Y.: Mining social images with distance metric learning for automated image tagging. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 197–206. ACM (2011)
8. Li, X., Snoek, C.G., Worring, M.: Learning social tag relevance by neighbor voting. *IEEE Trans. Multimedia* **11**(7), 1310–1322 (2009)
9. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 461–470. ACM (2010)
10. Ma, H., Zhu, J., Lyu, M.R.T., King, I.: Bridging the semantic gap between image contents and tags. *IEEE Trans. Multimedia* **12**(5), 462–473 (2010)
11. Gao, Y., Wang, M., Zha, Z.J., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. *IEEE Trans. Image Process.* **22**(1), 363–376 (2013)
12. Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C.G., Bimbo, A.D.: Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval. *ACM Comput. Surv. (CSUR)* **49**(1), 14 (2016)
13. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(3), 453–465 (2014)
14. Rohrbach, M., Stark, M., Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1641–1648. IEEE (2011)
15. Chao, W.-L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 52–68. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46475-6\\_4](https://doi.org/10.1007/978-3-319-46475-6_4)
16. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2927–2936 (2015)

17. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. arXiv preprint [arXiv:1707.00600](https://arxiv.org/abs/1707.00600) (2017)
18. Fu, Y., Xiang, T., Jiang, Y.G., Xue, X., Sigal, L., Gong, S.: Recent advances in zero-shot recognition: toward data-efficient understanding of visual content. *IEEE Signal Process. Mag.* **35**(1), 112–125 (2018)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
21. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(7), 1425–1438 (2016)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)
24. Loeff, N., Alm, C.O., Forsyth, D.A.: Discriminating image senses by clustering with multimodal features. In: *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pp. 547–554. Association for Computational Linguistics (2006)
25. Lazaridou, A., Bruni, E., Baroni, M.: Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1403–1414 (2014)
26. Lazaridou, A., Pham, N.T., Baroni, M.: Combining language and vision with a multimodal skip-gram model. arXiv preprint [arXiv:1501.02598](https://arxiv.org/abs/1501.02598) (2015)
27. Silberer, C., Ferrari, V., Lapata, M.: Visually grounded meaning representations. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(11), 2284–2297 (2017)
28. Xia, F., Liu, T.Y., Wang, J., Zhang, W., Li, H.: Listwise approach to learning to rank: theory and algorithm. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 1192–1199. ACM (2008)
29. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon’s Mechanical Turk. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pp. 139–147. Association for Computational Linguistics (2010)
30. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010)
31. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
32. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint [arXiv:1607.04606](https://arxiv.org/abs/1607.04606) (2016)
33. Honnibal, M., Montani, I.: spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017, to appear)
34. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. arXiv preprint [arXiv:1511.05879](https://arxiv.org/abs/1511.05879) (2015)
35. Paszke, A., et al.: Automatic differentiation in PyTorch. In: *NIPS-W* (2017)