

Original citation:

Chowdhury, Sreyasi Nag, Tandon, Niket, Ferhatosmanoglu, Hakan and Weikum, Gerhard (2017) VISIR : visual and semantic image label refinement. In: 11th ACM International Conference on Web Search and Data Mining (WSDM) 2018, Marina Del Rey, CA, USA, 5–9 Feb 2018. Published in: Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM) 2018

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/95881>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-5581-0/18/02. . . \$15.00
<https://doi.org/10.1145/3159652.3159693>

A note on versions:

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

VISIR: Visual and Semantic Image Label Refinement

Sreyasi Nag Chowdhury*

sreyasi@mpi-inf.mpg.de

*Max Planck Institute for Informatics

Niket Tandon†

nikett@allenai.org

† Allen Institute for Artificial Intelligence

Hakan Ferhatosmanoglu‡

hakan.f@warwick.ac.uk

‡ University of Warwick

Gerhard Weikum*

weikum@mpi-inf.mpg.de

*Max Planck Institute for Informatics

ABSTRACT

The social media explosion has populated the Internet with a wealth of images. There are two existing paradigms for image retrieval: 1) content-based image retrieval (CBIR), which has traditionally used visual features for similarity search (e.g., SIFT features), and 2) tag-based image retrieval (TBIR), which has relied on user tagging (e.g., Flickr tags). CBIR now gains semantic expressiveness by advances in deep-learning-based detection of visual labels. TBIR benefits from query-and-click logs to automatically infer more informative labels. However, learning-based tagging still yields noisy labels and is restricted to concrete objects, missing out on generalizations and abstractions. Click-based tagging is limited to terms that appear in the textual context of an image or in queries that lead to a click. This paper addresses the above limitations by semantically refining and expanding the labels suggested by learning-based object detection. We consider the semantic coherence between the labels for different objects, leverage lexical and commonsense knowledge, and cast the label assignment into a constrained optimization problem solved by an integer linear program. Experiments show that our method, called VISIR, improves the quality of the state-of-the-art visual labeling tools like LSDA and YOLO.

CCS CONCEPTS

• Information systems → World Wide Web; Multimedia content creation; Content analysis and feature selection;

KEYWORDS

Image Tagging, Semantic Coherence, Background Knowledge

1 INTRODUCTION

Motivation and Problem: The enormous growth of social media has populated the Internet with a wealth of images. On one hand, this makes image search easier, as there is redundancy for many keywords with informative text surrounding the images. On the other hand, it makes search harder, as there is a huge amount of visual contents that is hardly understood by the search engine. There are two paradigms for searching images: content-based image retrieval (CBIR) and tag-based image retrieval (TBIR).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5581-0/18/02...\$15.00

<https://doi.org/10.1145/3159652.3159693>



Figure 1: Noisy and Incomplete Labels: a) from LSDA [16] - dog, Browning machine gun, greater kudu, bird b) from flickr.com - happiness

CBIR finds images similar to a query image based on visual features that are used to represent an image. These features include color, shape, texture, SIFT descriptors etc. (e.g., [2, 7, 20]). Recent advances in deep-learning-based object detection have lifted this

	LSDA Labels	VISIR labels
	allosaurus loggerhead turtle person bird	person guitar stringed instrument self-expression
	bone china stove WC, loo cup or mug	food processor bowl cup or mug utensil
	cucumber snake green mamba	snake reptile slithery poisonous
	racket person bathing cap tennis ball head cabbage	tennis bat individual play tennis tennis ball

Figure 2: Images with labels from LSDA and VISIR

approach to a higher level, by assigning object labels to bounding boxes (e.g., [16, 35, 37, 42]). However, these labels are limited to concrete object classes (e.g., *truck, SUV, Toyota Yaris Hybrid 2016*, etc.), often trained (only) on (subsets of) the ca. 20,000 classes of ImageNet [8]. Thus, they miss out on generalizations (e.g., *vehicle*) and abstractions (e.g., *transportation, traffic jam, rush hour*). Fig 1 (a) shows the top-confidence visual labels by LSDA [16] for an example case of incorrect labels.

TBIR retrieves images by textual matches between user query and manually assigned image tags (e.g., from collaborative communities such as Flickr). While some of the semantic gap in CBIR is reduced in TBIR, the performance of TBIR often suffers from incomplete and ambiguous tags [15]. Figure 1 (b) illustrates this point: there is only a single tag *happiness* and none for the concrete objects in the image. For the big search engines, one way of overcoming this bottleneck is to exploit query-and-click logs (e.g., [6, 17, 45]). The query keyword(s) associated with a click can be treated as label(s) for the clicked image. However, this method crucially relies on the labels to appear in (sufficiently many) queries (or, traditionally, salient text surrounding the image). [22] gives a survey on TBIR and tag assignment and refinement.

Recently, the gap between the two image search paradigms is narrowing. TBIR-style tags inferred from query-and-click logs can be used to train a deep-learning network for more informative labels towards better CBIR. Also, crowdsourcing could be a way towards more semantic labels (e.g., [19]), for example, to capture human activities or emotions (e.g., [9, 13, 18, 33]). Nevertheless, there are still major shortcomings in the state-of-the-art.

This paper addresses the outlined limitations. The goal is to automatically annotate images with semantically informative tags, including generalizations and abstractions and also cleaning out noisy labels for concrete objects.

Approach and Contribution: We leverage state-of-the-art CBIR by considering the visual tags of an existing object detection tool (viz. LSDA [16] in our experiments) as a starting point. Note that there are multiple labels for each bounding box with varying confidence scores, and our goal is to compute the most informative labels for the entire image. We impose a constrained optimization on these initial labels, in order to enforce their semantic coherence. We also consider labels that are visually similar to the detected ones, to compensate for omissions. In addition, we utilize lexical and commonsense knowledge to generate candidate labels for generalizations (viz. hypernyms from WordNet [30]) and abstractions (viz. concepts from ConceptNet [40]). So we both refine and expand the initial labels. The joint inference on the entire label space is modeled as an optimization problem, solved by an integer linear program (viz. using the Gurobi ILP solver). Figure 2 shows examples for the input labels from the deep-learning-based visual object detection (left column) and the output labels that VISIR computes (right column). The labels from LSDA illustrate a clear semantic incoherence for these specific examples. VISIR labels are coherent, adds generalizations (in blue) and abstractions (in green). Incorrect labels are marked red.

Although our work aligns with the existing TBIR research on social tagging and tag refinement, there are key differences.

Granularity: Our starting point is labels for bounding boxes, whereas

user-provided tags refer to an entire image.

Cardinality: The number of bounding boxes in one image can be quite large. Moreover, object detectors usually produce a long list of varying-confidence labels for each bounding box.

Noise: As a result, many of the visual candidate labels in our approach are of mixed quality, whereas traditional social tagging typically has few but trusted annotations per image.

For these factors, our notion of tag refinement is unlike the one in prior work. Therefore, we refer to our task as *Visual Tag Refinement*.

Visual Tag Refinement can be broken down into three sub-tasks, for which this paper provides effective solutions: 1) elimination of incoherent tags¹ among the initial visual labels, 2) expansion of the tag space by adding visually similar tags missed by object detectors, and adding candidate tags for generalization and abstraction, 3) joint inference on the enriched tag space, by ILP.

2 RELATED WORK

Automatic Image Annotation: Early work on this problem generated tags only for an entire image (or a single image region), but did so one class at a time (e.g., [1]). More recent methods support labeling multiple objects in the same image. One such approach, WSABIE [44], performs k-nearest-neighbor classification on embeddings of words and features to scale to many classes.

State-of-the-art work on object detection addresses both recognizing object bounding boxes and tagging them with their class labels. Such work makes heavy use of deep learning (especially convolutional neural networks or CNNs). Prominent representatives are LSDA [16], and Faster R-CNN. The latter [37] improves the speed of object detection by incorporating a Region Proposal Network (RPN). Major emphasis in this line of Computer Vision work has been on coping with small, partly occluded and poorly illuminated objects. In contrast, the emphasis of VISIR is on the semantic coherence between objects and jointly modeling the uncertainty of candidate labels. Instead of speed, we optimize for higher labeling quality. VISIR is agnostic of the underlying object detector; it is straightforward to plug in a different tool.

Context plays an important role in computer vision [10], and context-based object detectors were popular before the success of CNNs. These methods consider local context [3], global context [32] or their combination [48]. With the advent of CNNs, the focus shifted to representation learning and improving detection speed. However, contextual information is gaining renewed attention. The state-of-the-art method [43] is based on a CNN-RNN combination, where the recurrent neural network (RNN) captures label dependencies. YOLO [35, 36] unifies learning with global context into a single neural network for the entire image. It exploits a word tree derived from the WordNet whose leaf nodes appear in ImageNet. In our experiments, we use YOLO as a baseline for context-aware object detection. All learning-based methods crucially rely on extensive training data.

Social Tagging: TBIR has its origin in community-based social tagging of images (e.g., Flickr), web pages or publications (e.g., Bibsonomy). Crowdsourcing to compile large training collections can be seen as a variant of this kind of user-provided tagging. There is ample research in this area [12, 21, 28], especially on learning tag

¹“tag” and “label” are used interchangeably in the paper with the same meaning.

recommendations. Our task of visual tag refinement differs from social tagging substantially. The CBIR-based tags that we start with, label individual objects instead of assigning tags to the overall image. Multiple candidate tags per bounding box also lead to a dense tag space in contrast to sparse user tags. Finally, the large number of varying-granularity and varying-confidence tags per image entails a much higher degree of noise in the label space, whereas social tags are usually considered trusted.

Tag Refinement: The problem of tag refinement aims at removing noisy tags from images while adding more relevant ones [22]. This line of work appears in the literature also as tag completion [11, 46] or image re-tagging [24, 25]. Most of this work uses only nouns as tags and disregards word ambiguity. Background knowledge, such as WordNet synonymy sets and other lexical relations, is rarely used. Word categories that are vital for denoting abstractions, namely, adjectives, verbs and verbal phrases, are out of scope. Moreover a common assumption is that visually similar images are semantically similar, meaning that they should have similar tags. This assumption is often invalid. This body of work employs a variety of methods, including metric learning [14], matrix completion [11, 46, 50], latent topic models [47], and more.

Commonsense Knowledge for Image Retrieval: The first work on image retrieval with commonsense knowledge (CSK) [26] exploited the Open Mind Commonsense Knowledgebase [39], a small knowledge base with simple properties of concepts for concept expansion and for activation spreading. Since then, much more comprehensive CSK knowledge bases have been constructed, most notably, ConceptNet [40] and WebChild [41]. However, such background knowledge has not been used by modern object detectors. A notable exception addresses emotions invoked by images, and tags objects with sentiment-bearing adjectives [4]. However, this work is limited to a small label space. A recent framework for image search [5] uses CSK extracted through OpenIE for query expansion.

3 MODEL AND METHODOLOGY

We define the problem of *Visual Tag Refinement* as the tasks of –

- 1) cleaning noisy object tags from low-level image features
- 2) enriching existing detections by adding additional relevant tags
- 3) abstracting from concrete objects towards a more conceptual space.

We first present a framework for the proposed problem, followed by the description of its individual components. We then present the optimization model to solve the problem.

3.1 Framework for Visual Tag Refinement

We consider an image x with multiple bounding boxes x_1, x_2, \dots, x_k . Each bounding box x_i has labels for detected physical objects along with detection confidence scores. The values of these labels and scores are outputs of an off-the-shelf object detection tool, e.g., LSDA [16]. We define three different label spaces, candidates from which would be associated either with bounding boxes of an image or would globally add semantics to it:

- A space of all possible object labels detectable by an underlying object detection tool is denoted by CL (for Concrete-object Labels). For example, the LSDA tool [16] uses ImageNet [8] object classes, which are also leaf-level synsets of WordNet [30]. There can be two sets of such classes for a bounding box x_i of

image x : cl_i constitute those labels originally detected from low-level image features, cl'_i constitute undetected labels visually similar to those in cl_i .

- A space of extended labels is denoted by XL . For each image x and bounding box x_i , a subset from this space, xl_i , contains additional label candidates that generalize the classes in cl_i and cl'_i . For example, “ant” $\in cl_i \rightarrow$ “insect” $\in xl_i$. Adding generalized terms to the label space serves a dual purpose – overcoming the training bias of the object detection tools, and broadening the label space for greater web visibility. We discuss more on the issue of training bias in section 6.

CL and XL contain labels signifying visual objects in an image. Hence we call the super set of such labels “visual labels” VL ; $VL = CL \cup XL$.

- A space of abstract labels (utilities, emotions, themes) is denoted by AL . This constitutes abstract concepts associated with visual objects derived from commonsense knowledge bases. For example, $fragrant \in AL$ from the ConceptNet [27] clause `hasProperty(flower , fragrant)`.

An image x can hence be described by three sets of labels – a set of deep-learning based class labels $cl \cup cl' \in CL$, a set of extended labels $xl \in XL$, and a set of abstract labels $al \in AL$. Further, we define three different scores that act as edges between nodes of the above spaces:

- Visual Similarity $vsim(l_j, l_k)$ for $l_j, l_k \in CL$
- Semantic Relatedness $srel(l_j, l_k)$ for $l_j, l_k \in VL$
- Abstraction Confidence $aconf(l_j, al_n)$ for $al_n \in AL$ and $l_j \in VL$

We present the visual tag refinement problem in terms of three sub-problems:

- The noisy tag problem – for each image x and bounding box x_i infer which of the labels in $cl_i \in CL$ should be accepted. We eliminate those labels which are not coherent with the other bounding box detections in the image. For example, in Figure 2 image 5 we eliminate the detection *cucumber* since it is not semantically related to the other labels *snake* and *green mamba*.
- The incomplete tag problem – for each image x and bounding box x_i infer which of the labels in $cl'_i \in CL$ and $xl_i \in XL$ should be additionally associated with the bounding box.
- The abstraction tag problem – infer which of $al \in AL$ should be globally associated with image x .

We solve these problems jointly and retain the most confident hypothesis for each bounding box relative to the others as well as a global hypothesis toward tag abstraction in an image. Hence, we predict a set of plausible labels $L_x \in CL \cup XL \cup AL$ for an image x .

3.2 Visual Similarity (or “Confusability”)

Deep-learning based tools using low-level image features to predict the object classes can confuse one object to be another. We consider two labels to be visually similar if they occur as candidates in cl_i for the same bounding box x_i . We collect evidence of such visual similarity from low-level image features, in particular, from object detection results of LSDA [16]. We define the visual similarity between two labels l_j and l_k by a Jaccard-style similarity measure

as shown in Equation 1. In this similarity measure, if labels l_j and l_k always appear together as candidates for the same bounding box, and never with any other labels, then they are considered highly visually similar, $vsim(l_j, l_k) = 1$. If labels l_j and l_k never appear together, one label is never confused by the tool to be another; in this case $vsim(l_j, l_k) = 0$, meaning l_j and l_k are not visually similar. Given that the initial object detections from low level image features are noisy in itself, this evidence would also contain noise. However, it is expected that the evidence will hold when it is computed over a large dataset.

$$vsim(l_j, l_k) = \frac{\sum_{i: l_j, l_k \in c_i} (conf_{BB}(x_i, l_j) + conf_{BB}(x_i, l_k))}{\sum_{i: l_j \in c_i} (conf_{BB}(x_i, l_j)) + \sum_{i: l_k \in c_i} (conf_{BB}(x_i, l_k))} \quad (1)$$

We can refer to this measure also as “confusability” since the object detection tool confuses one object to be another based on similar low-level visual features.

3.3 Semantic Relatedness

Semantic Relatedness between two concepts signifies their conceptual similarity. Our model uses this measure to establish the contextual coherence between labels of different bounding boxes. The relatedness between two labels l_j and l_k is defined as a weighted linear combination of their cosine similarity from word embeddings and their spatial co-location confidence.

$$srel(l_j, l_k) = \delta cosine(l_j, l_k) + (1 - \delta) coloc(l_j, l_k) \quad \text{for } l_j, l_k \in VL \quad (2)$$

Word Embeddings: To improve the contextual coherence between object labels in images, the context of words needs to be captured. We utilize vector space word embeddings for this purpose. A word2vec [29] model is trained from manually annotated image descriptions from a large set of image captions, as described later in more details. The cosine similarity between two labels – $cosine(., .)$ in Equation 2 – is calculated from their respective word vectors.

Spatial Co-location: Spatial relationships between concepts carry an important evidence of relatedness. For example, an “apple” and a “table” are related concepts since they occur in close spatial proximity. Similarly, a “tennis racket” and a “lemon” are unrelated. $coloc(., .)$ in Equation 2 is a frequency-based co-location score mined from manual annotations of image labels.

3.4 Concept Generalization

A hypernym is a superordinate of a concept. In other words, a concept is an instantiation of its hypernym. For example, *fruit* is a hypernym for *apple*, i.e., *apple* IsA *fruit*. WordNet [30] provides a hierarchy of concepts and their hypernyms which we leverage to generalize our object classes. WordNet also reports different meanings (senses) of a concept; for example a *punching bag* is (*a person on whom another person vents their anger*) or (*an inflated ball or bag that is suspended and punched for training in boxing*), leading to very different hypernymy trees. For this reason, we map our object classes from ImageNet into their correct WordNet sense number, followed by traversing their hypernymy tree up to a certain

level. This yields a cleaner generalization. Further more, to avoid exotic words among the hypernyms, we use their approximate Google result counts and prune out those below a threshold. Hence for the concept *ant* we retain the hypernym *insect* and prune the hypernym *hymenopteran*. Following this heuristics, we assign 1 to 3 hypernyms per object class.

3.5 Concept Abstraction

To introduce human factors like commonsense and subjective perception, we incorporate abstract words and phrases associated with visual concepts of an image. For example an *accordion* is “used to” *make music*. We consider two relations from ConceptNet 5 [27] for assigning the abstract labels - *usedFor*, and *hasProperty*. Some example of assigned abstract labels/phrases (in green) can be found in Figure 2. Abstract concepts which are assigned to images have high abstraction confidence. Abstraction confidence of a concept/phrase is defined as the joint semantic relatedness of the phrase and the refined visual labels of the image.

3.6 Tag Refinement Modeled as an ILP

We cast the multi-label visual tag refinement problem into an Integer Linear Program (ILP) optimization with the following definitions. We choose ILP as it is a very expressive framework for modeling constrained optimization (more powerful than probabilistic graphical models), and at the same time comes with very mature and efficient solvers like Gurobi (<http://gurobi.com>). Some tools for probabilistic graphical models even use ILP for efficient MAP inference.

Given an image x , with bounding boxes x_1, x_2, \dots , it has three sets of visual labels: cl_i (initial bounding box labels), cl'_i (labels visually similar to the original detections), and xl_i (hypernyms of labels in $cl_i \cup cl'_i$). The set $vl_i = cl_i \cup cl'_i \cup xl_i$ constitutes all visual labels which are candidates for bounding box x_i . The image would also be assigned abstract labels al_1, al_2, \dots globally. We thus introduce 0-1 decision variables:

$X_{ij} = 1$ if x_i should indeed have visual label vl_j , 0 otherwise

$Y_j = 1$ if x should indeed have abstract label al_j , 0 otherwise

$Z_{ijmk} = 1$ if $X_{ij} = 1$ and $X_{mk} = 1$, 0 otherwise

$W_{ijk} = 1$ if $X_{ij} = 1$ and $Y_k = 1$, 0 otherwise

Decision variables Z_{ijmk} and W_{ijk} emphasise pair-wise coherence between two visual labels and between a visual and an abstract label respectively.

Objective: Select labels for x and its bounding boxes which maximizes a weighted sum of evidence and coherence –

$$\max \left[\alpha \sum_{i,j} (vconf(x_i, l_j) + \kappa gconf(x_i, l_j)) X_{ij} + \beta \sum_{i,m} \sum_{\substack{l_j \in vl_i \\ l_k \in vl_m}} srel(l_j, l_k) Z_{ijmk} + \gamma \sum_{l_j \in VL} \sum_k aconf(l_j, al_k) \sum_i W_{ijk} \right] \quad (3)$$

with hyper-parameters $\alpha, \beta, \gamma, \kappa$.

For each $l \in CL$, we define set $S(l) \subseteq CL$ of labels visually similar to l . $vsim(l, l') = 0$ if $l' \notin S(l)$. Recall the definition of $vsim(., .)$ from Equation 1.

Visual Confidence, the confidence with which a visual label should be associated with an image is defined as:

$$\begin{aligned} vconf(x_i, l_j) &= conf_{BB}(x_i, l_j) \text{ if } l_j \in cl_i & (4) \\ &= \sum_{l \in cl_i} conf_{BB}(x_i, l) vsim(l, l_j) \text{ if } l_j \in cl'_i/cl_i & (5) \end{aligned}$$

Here, a high confident original detection adds significant weight to the objective function, hence increasing the chances of its retention. Similarly, the weight of a label visually similar to multiple original labels is boosted. Also, labels visually similar to only one low confident original label is assigned less importance.

For $l \in CL$ we define a set $H(l) \in XL$ of hypernyms of l . The *Generalization Confidence* of a label l_j in bounding box x_i is defined in terms of the semantic relatedness between the label and its hypernym.

$$\begin{aligned} gconf(x_i, l_j) &= \sum_{l: l_j \in H(l)} srel(l_j, l) \text{ if } l_j \in xl_i & (6) \\ &= 0 \text{ if } l_j \in \{cl_i \cup cl'_i\} & (7) \end{aligned}$$

Abstraction Confidence $aconf(., .)$ of a label l_j and an abstract concept al_k is defined as their semantic relatedness, weighted by the score of the assertion containing the abstract concept in ConceptNet [27]. For example, `hasProperty(baby, newborn)` has a score of 10.17 in ConceptNet. We name this score $CNet(al_k)$.

$$aconf(l_j, al_k) = CNet(al_k) * srel(l_j, al_k) \quad (8)$$

Constraints:

$\sum_j X_{ij} \leq 1$: for each bounding box x_i there can be at most one visual label ($\in VL$)

$\sum_j Y_j \leq 5$: one image x can have at most five abstract labels ($\in AL$)

$$\left. \begin{aligned} (1 - Z_{ijmk}) &\leq (1 - X_{ij}) + (1 - X_{mk}) \\ Z_{ijmk} &\leq X_{ij} \\ Z_{ijmk} &\leq X_{mk} \end{aligned} \right\} \begin{array}{l} \text{Pair-wise mutual} \\ \text{coherence between} \\ \text{visual labels} \end{array}$$

$$\left. \begin{aligned} (1 - W_{ijk}) &\leq (1 - X_{ij}) + (1 - Y_k) \\ W_{ijk} &\leq X_{ij} \\ W_{ijk} &\leq Y_k \end{aligned} \right\} \begin{array}{l} \text{Pair-wise mutual} \\ \text{coherence between} \\ \text{visual and abstract} \\ \text{labels} \end{array}$$

The final set of visual and abstract labels per image are expected to be highly coherent. This is validated in Section 5.

4 DATA SETS AND TOOLS

In this section, we present the image data sets as well as the criteria and heuristics we follow to mine the various background knowledge utilized in our optimization model.

ImageNet Object Classes: LSDA [16] is used to get the initial visual object labels from low-level image features. The LSDA tool has been trained on 7604 leaf-level nodes of ImageNet [8]. Most of these object classes are exotic concepts which rarely occur in everyday images. Examples include scientific names of flora and fauna – *interior live oak*, *Quercus wislizenii*, *American white oak*,

Quercus alba, and obscure terms – *pannikin*, *reliquary*, *lacrosse*. We prune those exotic classes by thresholding on their Google and Flickr search result counts. Some object class names are ambiguous where two senses of the same word from WordNet have been included. We consider only the most common sense. We work with the most frequent 1000 object classes obtained after pruning².

WordNet Hypernyms: For the ImageNet object classes described above, we traverse the WordNet hypernymy tree of the associated sense up to level three. We restrict the traversal level to avoid too much generalization – for example, *person* generalizing to *organism*. We prune out hypernyms with Google and Flickr result counts below a threshold. By considering the hypernyms of the 1000 ImageNet object classes mentioned above, we add 800 new visual labels to the model.

The ImageNet object classes and the WordNet hypernyms together constitute the **Visual Labels** of VISIR.

Abstract Labels: Commonsense knowledge (CSK) assertions from ConceptNet [40] contribute to concept abstraction in VISIR. For example, in Figure 2, the abstract concept *poisonous* is added to the labels of the fifth image. ConceptNet is a crowd-sourced knowledge base where most assertions have the default confidence score of 1.0 (as they were stated only by one person). Only popular statements like `hasProperty(apple, red fruit)` are stated by multiple people, hence raising the confidence score significantly. Certain assertions have contradictory scores – for example, `usedFor(acne medicine, clear skin)` appears twice, with scores 1.0 and -1.0. This happens when someone down-votes a statement. Using such indistinctive scores in VISIR would be uninformative. We therefore use the joint semantic relatedness of the assertion and visual labels of an image, weighted by the ConceptNet score (only positive scores), as the abstraction confidence.

Visual Similarity: The visual similarity or “confusability” scores (Equation 1) are mined from object detection results (from low-level image features) over 1 million images from the following data sets that are popularly employed in the computer vision community: Flickr 30k [49], Pascal Sentence Dataset [34], SBU Image Dataset [31], MS-COCO [23]. All these data sets have collections of Flickr images not pertaining to any particular domain. For each detected bounding box, LSDA provides a confidence score distribution over 7604 object classes (leaf nodes in ImageNet). Only predictions with a positive confidence score are considered as candidates for a bounding box. An object class pair appearing as candidates for the same bounding box are considered as visually similar. Table 1 shows few examples of visually similar object class pairs – *mail train* and *commuter train* are confused 91% times whereas *diaper* and *plaster cast* are confused 18% times.

Spatial Co-location: Spatial co-location scores between different object classes are mined from ground truth annotations of the detection challenge (DET) of ImageNet ILSVRC 2015 [38]. We consider two objects to be spatially co-located only if they are tagged in the same image. For simplicity, we do not consider the physical distance between the bounding boxes of the tagged object classes.

²The full list is available at http://people.mpi-inf.mpg.de/~sreyasi/visTagRef/1000classes_names.txt

Table 1: Object class pairs and visual similarity scores

object1	object2	visual similarity
mail train	commuter train	0.91
cattle	horse	0.76
soccer ball	kite balloon	0.26
Red Delicious	bowling ball	0.21
diaper	plaster cast	0.18
bicycle pump	mascara	0.17

A frequency-based co-location score is assigned to pairs of object classes based on evidence over the train set of ILSVRC DET. We find spatial co-location data for 200 object classes (since the detection challenge only considers 200 object classes). The top few frequently co-located objects are: *(person, microphone)*, *(table, chair)*, *(person, sunglasses)*, *(person, table)*, *(person, chair)*. A general observation would be that the image collection in ILSVRC DET has a high occurrence of *person*.

5 EXPERIMENTS AND RESULTS

We analyze and compare the results that VISIR produces with that of two baselines: LSDA³ and YOLO⁴. The performances of LSDA, YOLO, and VISIR are compared on the basis of precision, recall and F1-score measures.

5.1 Setup

As discussed in Section 3.1, we operate with three kinds of labels: visual class labels from ImageNet (*CL*), their generalizations (*XL*) which consist of WordNet hypernyms of labels $\in CL$, and abstract labels (*AL*) from commonsense knowledge. We evaluate the methods with respect to three different label spaces (as the combination of three types of labels): *CL*, *CL + XL*, and *CL + XL + AL*. LSDA and YOLO operate only on *CL*, while VISIR has three variants (configuring it for the above combinations of label spaces). Each system is given a label budget of 5 tags per image. For VISIR, this is enforced by an ILP constraint; for the two baselines, we use their confidence scores to pick the top-5.

Hyper-parameter Tuning: To tune the hyper-parameters for Equation 3 we use the annotations of the training image set of ILSVRC DET. We also extend this set by adding the hypernyms of the ground-truth labels as correct labels. A randomized search is used to tune the hyperparameters.

User Evaluation: Besides establishing semantic coherence among concrete object labels, VISIR applies concept generalization and abstraction. For modern benchmark datasets like ILSVRC 2015 DET, such enriched labeling does not exist so far. Therefore, in order to evaluate VISIR and compare to baselines, we construct a labeled image dataset by collecting human judgments about correctness of labels as discussed below.

For each label space, *CL*, *CL + XL*, and *CL + XL + AL*, the union of the labels produced by each method forms the set of result labels for an image. This result pool is evaluated by human annotators. Judges determined whether each label is appropriate and informative for

an image. Instead of a binary assessments, annotators are asked to grade each label in the pool with 0, 1, or 2 – 0 corresponding to incorrect labels, 2 corresponding to highly relevant labels. We gather user judgments for the three label pools (corresponding to the label spaces) separately. This produces three different sets of graded labels per image. Users are not informed about the nature of the label pools. For each label pool we collect responses from at least 5 judges. The final assessment is determined by the majority of the judges (e.g., at least 3 out of 5 need to assert that a label is good).

Selection of Test Images: A major goal of this work is to make the refined labels more coherent or semantically related. Hence, we focus on the case where the deep-learning-based object detection tools produce contextually incoherent results. For the user evaluation, we collect a set of images with a reasonable context – those that have 3-7 detected bounding boxes and with LSDA labels having a semantic relatedness score less than 0.1. Such 100 images are collected from the ILSVRC 2015 DET Val image set.

5.2 Model Performance

Precision is estimated as the fraction of “good labels” detected, where a “good label” is one considered relevant by the majority of the human judges. We assess the recall per method as the number of labels picked from the good labels in the pool of labels generated by all three methods. The recall is artificially restricted because the label pool may contain more good labels than the label budget of the method. For example, if the label budget per method is set to 5, even if all 5 labels of a method are good, the recall for a pool with 8 good labels would only be 5/8. However, it is a fair notion across the different methods.

Relaxed vs Conservative Assessments: According to the evaluation design, labels graded 1 are either inconspicuous, or less relevant to the image than labels graded 2. In order to identify the “good labels” in a label pool, we define two methods of assessment: *Relaxed Assessment* considers all labels graded 1 or 2 as correct. *Conservative Assessment* considers only those labels graded 2 as correct, resulting in a stricter setup. The three graded label pools from the user evaluations naturally have labels in common.

Performance Results: Tables 2 through 4 compare the three methods for the three different label pools – *CL*, *CL + XL*, *CL + XL + AL* – with conservative assessment. For *CL*, there is no real improvement over LSDA, but we see that for *CL + XL* and *CL + XL + AL* VISIR adds a good number of semantically informative labels and improves on the two baselines in terms of both precision and recall.

We also test VISIR’s performance with a tighter constraint on choosing the number of bounding boxes per image, by setting the label budget to 80% of all bounding boxes received as input. This variant, which we refer to as VISIR*, aims to filter out more noise in the output of the deep-learning-based object detections. Naturally, VISIR*-CL would have higher precision than VISIR-CL while sacrificing on recall. VISIR*-CL improves further on precision and F1-score because it is able to eliminate some of the initial noise the LSDA detections bring in. For pools *CL + XL* and *CL + XL + AL*, VISIR* has higher precision than VISIR, but slightly loses in recall.

³<http://lsda.berkeleyvision.org/>

⁴<https://pjreddie.com/darknet/yolo/>

Table 2: Pool CL: Conservative Assessment

System	Precision	Recall	F1-score
LSDA	0.51	0.86	0.64
YOLO	0.49	0.56	0.52
VISIR-CL	0.51	0.86	0.64
VISIR [*] -CL	0.57	0.81	0.67

Table 3: Pool CL+XL: Conservative Assessment

System	Precision	Recall	F1-score
LSDA	0.52	0.81	0.63
YOLO	0.49	0.51	0.50
VISIR-CL+XL	0.54	0.82	0.65
VISIR [*] -CL+XL	0.60	0.76	0.67

Table 4: Pool CL+XL+AL: Conservative Assessment

System	Precision	Recall	F1-score
LSDA	0.49	0.35	0.41
YOLO	0.52	0.23	0.32
VISIR-CL+XL+AL	0.54	0.91	0.68
VISIR [*] -CL+XL+AL	0.56	0.89	0.69

Table 5 and Table 6 show the relaxed and conservative assessments with respect to the combined pool (i.e., for all three label spaces together) of good labels per image. It is natural that all methods perform better for the relaxed setting compared to that of the conservative assessment. However, the observation that VISIR’s performance does not degrade much for the conservative assessment demonstrates its high output quality and robustness. Figure 3 illustrates this by anecdotal examples with the labels assigned by each of the competitors (with good labels in black and bad ones in red). In image 4, LSDA produces typically unrelated labels – a *monkey* and a *tennis ball*. This contextual incoherence likely arises due to low level color features. In contrast to LSDA, YOLO addresses the semantic coherence of the labels, however likely in expense of recall (for example in image 6). By necessitating semantic coherence among detected labels VISIR eliminates incoherent labels - for example, VISIR removes *motorcycle* from image 1, *tennis ball* from image 4, *hat with a wide brim* from image 5 and so on.

Table 5: Aggregate Pool: Relaxed Assessment

System	Precision	Recall	F1-score
LSDA	0.55	0.30	0.39
YOLO	0.57	0.19	0.29
VISIR-CL	0.57	0.28	0.38
VISIR-CL+XL	0.62	0.30	0.40
VISIR-CL+XL+AL	0.71	0.90	0.79

Table 7 lists the new labels introduced by VISIR, each for at least 10 images. These labels are generated via generalization (from WordNet hypernyms) and abstraction (from commonsense knowledge). As none of the baselines can produce these labels, VISIR naturally achieves a recall of 1. The precision values for the labels illustrate how VISIR addresses the problem of label incompleteness. In most cases, these labels were assessed as correct by the judges.

Table 6: Aggregate Pool: Conservative Assessment

System	Precision	Recall	F1-score
LSDA	0.49	0.35	0.41
YOLO	0.52	0.23	0.32
VISIR-CL	0.52	0.34	0.41
VISIR-CL+XL	0.55	0.35	0.43
VISIR-CL+XL+AL	0.54	0.91	0.68

Table 7: New labels suggested by VISIR

Label	Label frequency	Precision
individual	46	0.59
man or woman	44	0.64
animal	31	0.94
human	20	0.95
canine	18	0.94
furniture	12	0.83
barking animal	11	1.00

6 DISCUSSION OF LIMITATIONS

Training Bias in LSDA: The LSDA tool predicts only leaf-level object classes of ImageNet. The same limitation holds for most other state-of-the-art object detectors. Because of this incomplete tag space many important objects cannot be detected. For example, *giraffe* is not a leaf-level object class of ImageNet. Since LSDA did not see any training images of a giraffe, it mis-labels objects in Figure 1a according to its training. This noise propagates to our model, sometimes making it impossible to find the correct labels.

Incorrect Sense Mapping in ImageNet: LSDA trains on ImageNet images. Hence improper word sense mappings in ImageNet propagate to incorrect labels from LSDA. For example, ImageNet contains similar images for two separate synsets *Sunglass (a convex lens used to start a fire)* and *Sunglasses (shades, dark glasses)*. Naturally, these two synsets have completely different WordNet hypernyms which VISIR uses, hence introducing noise. The direct hypernym of *Sunglass* is *lens*, while that of *Sunglasses* is *glasses*.

Incomplete Spatial Co-location data: Spatial co-location patterns mined from text contain noise due to linguistic variations in the form of proverbs. For example, the commonsense knowledge base WebChild [41] assigns significant confidence to the spatial co-location of *elephant* and *room* (most likely from the idiom “the elephant in the room”). To counter such linguistic bias, we have mined spatial co-location information from the manually annotated ground truth of ILSVRC [38]. Unfortunately, annotations are available for only 200 object classes, leaving us with only a small fraction of annotated visual-label pairs. If more cues of this kind were available, we would have been able to establish stronger contextual coherence.

Incomplete and Noisy Commonsense Knowledge: ConceptNet and WebChild are quite incomplete; so we cannot assign an abstract concept to every detected visual label. Also, assertions in these knowledge bases are often contradictory and noisy. We manage to reduce the noise by considering semantic relatedness with the visual labels, but this only alleviates part of the problem.

	LSDA	YOLO	VISIR-CL	VISIR-CL+XL	VISIR-CL+XL+AL
	person table motorcycle	bench person bowl	person table	person table	person table man or woman furniture
	table car	tow truck bench car chair	table chair	table chair	seat furniture chair flat dining furniture table
	monkey tennis ball	bird dog	monkey	monkey	primate monkey orangutan ape simian furry
	bird hat with a wide brim	airplane bird	bird	bird	bird avian flying animal
	helmet person watercraft smelling bottle bathing cap record sleeve impeller	person	bathing cap bib watercraft helmet person	bathing cap fabric watercraft helmet person	bathing cap cloth, fabric watercraft helmet protective hat individual person
	table baby bed swimming trunks	chair	chair table	chair table	seat furniture chair table flat

Figure 3: Images with labels from LSDA, YOLO, and different configurations of VISIR

7 CONCLUSION

We presented VISIR, a new method for refining and expanding visual labels for images. Its key strengths are cleaning out noisy labels from predictions by object detection tools and adding informative labels that capture generalizations and abstractions. Our model makes this feasible by considering the visual similarity of labels, the semantic coherence across concepts, and various kinds of background knowledge. The joint inference on an enriched label

candidate space is performed by means of a judiciously designed Integer Linear Program. Our experiments show the viability of the approach, and also demonstrate significant improvements over two state-of-the-art object detection and tagging tools.

ACKNOWLEDGEMENTS

This work was partly supported by the Alexander von Humboldt Foundation.

REFERENCES

- [1] Emre Akbas and Fatos T. Yarman-Vural. 2007. Automatic Image Annotation by Ensemble of Visual Descriptors. In *CVPR*. <https://doi.org/10.1109/CVPR.2007.383484>
- [2] Relja Arandjelovic and Andrew Zisserman. 2012. Three things everyone should know to improve object retrieval. In *CVPR*. <https://doi.org/10.1109/CVPR.2012.6248018>
- [3] João Carreira, Fuxin Li, and Cristian Sminchisescu. 2012. Object Recognition by Sequential Figure-Ground Ranking. *International Journal of Computer Vision* (2012). <https://doi.org/10.1007/s11263-011-0507-2>
- [4] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. 2014. Object-Based Visual Sentiment Content Analysis and Application. In *ACM MM*. <http://doi.acm.org/10.1145/2647868.2654935>
- [5] Sreyasi Nag Chowdhury, Niket Tandon, and Gerhard Weikum. 2016. Know2Look: Commonsense Knowledge for Visual Search. In *5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT*.
- [6] Nick Craswell and Martin Szummer. 2007. Random Walks on the Click Graph. In *SIGIR*. <http://doi.acm.org/10.1145/1277741.1277784>
- [7] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40, 2 (2008), 5:1–5:60. <http://doi.acm.org/10.1145/1348246.1348248>
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. <https://doi.org/10.1109/CVPRW.2009.5206848>
- [9] Santosh Kumar Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning Everything about Anything: Webly-Supervised Visual Concept Learning. In *CVPR*. <https://doi.org/10.1109/CVPR.2014.412>
- [10] Santosh Kumar Divvala, Derek Hoiem, James Hays, Alexei A. Efros, and Martial Hebert. 2009. An Empirical Study of Context in Object Detection. In *CVPR*. <https://doi.org/10.1109/CVPRW.2009.5206532>
- [11] Zheyun Feng, Songhe Feng, Rong Jin, and Anil K. Jain. 2014. Image Tag Completion by Noisy Matrix Recovery. In *ECCV*. https://doi.org/10.1007/978-3-319-10584-0_28
- [12] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. 2013. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. *IEEE Trans. Image Processing* 22, 1 (2013), 363–376. <https://doi.org/10.1109/TIP.2012.2202676>
- [13] Georgia Gkioxari, Ross B. Girshick, and Jitendra Malik. 2015. Contextual Action Recognition with R²CNN. In *ICCV*. <https://doi.org/10.1109/ICCV.2015.129>
- [14] Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cordelia Schmid. 2009. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*. <https://doi.org/10.1109/ICCV.2009.5459266>
- [15] Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The complex dynamics of collaborative tagging. In *WWW*. <http://doi.acm.org/10.1145/1242572.1242602>
- [16] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross B. Girshick, Trevor Darrell, and Kate Saenko. 2014. LSDA: Large Scale Detection through Adaptation. In *NIPS*. <http://papers.nips.cc/paper/5418-lsda-large-scale-detection-through-adaptation>
- [17] Xian-Sheng Hua and Jin Li. 2014. Tell me what. In *ICME Workshops*. <https://doi.org/10.1109/ICMEW.2014.6890616>
- [18] Ronak Kosti, Jose M. Alvarez, Adrià Recasens, and Àgata Lapedriza. 2017. Emotion Recognition in Context. In *CVPR*. <https://doi.org/10.1109/CVPR.2017.212>
- [19] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. 2016. Crowdsourcing in Computer Vision. *Foundations and Trends in Computer Graphics and Vision* 10, 3 (2016), 177–243. <https://doi.org/10.1561/06000000071>
- [20] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. 2006. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP* 2, 1 (2006), 1–19. <http://doi.acm.org/10.1145/1126004.1126005>
- [21] Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009. Learning Social Tag Relevance by Neighbor Voting. *IEEE Trans. Multimedia* 11, 7 (2009), 1310–1322. <https://doi.org/10.1109/TMM.2009.2030598>
- [22] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. 2016. Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval. *ACM Comput. Surv.* 49, 1 (2016), 14:1–14:39. <http://doi.acm.org/10.1145/2906152>
- [23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*. https://doi.org/10.1007/978-3-319-10602-1_48
- [24] Dong Liu, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. 2010. Retagging social images based on visual and semantic consistency. In *WWW*. <http://doi.acm.org/10.1145/1772690.1772848>
- [25] Dong Liu, Shuicheng Yan, Xian-Sheng Hua, and Hong-Jiang Zhang. 2011. Image Retagging Using Collaborative Tag Propagation. *IEEE Trans. Multimedia* 13, 4 (2011), 702–712. <https://doi.org/10.1109/TMM.2011.2134078>
- [26] Hugo Liu and Henry Lieberman. 2002. Robust photo retrieval using world semantics. In *Proceedings of LREC2002 Workshop: Using Semantics for IR*. 15–20.
- [27] Hugo Liu and Push Singh. 2004. ConceptNet – a practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.
- [28] Leandro Balby Marinho, Alexandros Nanopoulos, Lars Schmidt-Thieme, Robert Jäschke, Andreas Hotho, Gerd Stumme, and Panagiotis Symeonidis. 2011. Social Tagging Recommender Systems. In *Recommender Systems Handbook*. 615–644. https://doi.org/10.1007/978-0-387-85820-3_19
- [29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 3111–3119.
- [30] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41. <http://doi.acm.org/10.1145/219717.219748>
- [31] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*. <http://papers.nips.cc/paper/4470-im2text-describing-images-using-1-million-captioned-photographs>
- [32] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge J. Belongie. 2007. Objects in Context. In *ICCV*. <https://doi.org/10.1109/ICCV.2007.4408986>
- [33] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rosenberg, and Fei-Fei Li. 2015. Learning semantic relationships for better action retrieval in images. In *CVPR*. <https://doi.org/10.1109/CVPR.2015.7298713>
- [34] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon’s Mechanical Turk. In *Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. 139–147. <http://aclanthology.info/papers/W10-0721/collecting-image-annotations-using-amazon-s-mechanical-turk>
- [35] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*. <https://doi.org/10.1109/CVPR.2016.91>
- [36] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *CVPR*. <https://doi.org/10.1109/CVPR.2017.690>
- [37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.
- [38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- [39] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge Acquisition from the General Public. In *On the Move to Meaningful Internet Systems (Lecture Notes in Computer Science)*, Vol. 2519. Springer, 1223–1237. https://doi.org/10.1007/3-540-36124-3_77
- [40] Robert Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *LREC*. <http://www.lrec-conf.org/proceedings/lrec2012/summaries/1072.html>
- [41] Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. WebChild: harvesting and organizing commonsense knowledge from the web. In *WSDM*. <http://doi.acm.org/10.1145/2556195.2556245>
- [42] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 652–663. <https://doi.org/10.1109/TPAMI.2016.2587640>
- [43] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *CVPR*. <https://doi.org/10.1109/CVPR.2016.251>
- [44] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. WSABIE: Scaling Up to Large Vocabulary Image Annotation. In *IJCAI*.
- [45] Fei Wu, Xinyan Lu, Jun Song, Shuicheng Yan, Zhongfei (Mark) Zhang, Yong Rui, and Yueting Zhuang. 2016. Learning of Multimodal Representations With Random Walks on the Click Graph. *IEEE Trans. Image Processing* 25, 2 (2016), 630–642. <https://doi.org/10.1109/TIP.2015.2507401>
- [46] Lei Wu, Rong Jin, and Anil K. Jain. 2013. Tag Completion for Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 3 (2013), 716–727. <https://doi.org/10.1109/TPAMI.2012.124>
- [47] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. 2009. Tag refinement by regularized LDA. In *ACM MM*. <http://doi.acm.org/10.1145/1631272.1631359>
- [48] Jian Yao, Sanja Fidler, and Raquel Urtasun. 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*. <https://doi.org/10.1109/CVPR.2012.6247739>
- [49] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78. <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229>
- [50] Guanyu Zhu, Shuicheng Yan, and Yi Ma. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM MM*. <http://doi.acm.org/10.1145/1873951.1874028>