Picture Tags and World Knowledge

Learning Tag Relations from Visual Semantic Sources

Lexing Xie, Xuming He
Australian National University, NICTA
Canberra, Australia

ABSTRACT

This paper studies the use of everyday words to describe images. The common saying has it that *a picture is worth a thousand words*, here we ask *which thousand*? The proliferation of tagged social multimedia data presents a challenge to understanding collective tag-use at large scale – one can ask if patterns from photo tags help understand tag-tag relations, and how it can be leveraged to improve visual search and recognition. We propose a new method to jointly analyze three distinct visual knowledge resources: Flickr, ImageNet/WordNet, and ConceptNet. This allows us to quantify the visual relevance of both tags learn their relationships. We propose a novel network estimation algorithm, Inverse Concept Rank, to infer incomplete tag relationships. We then design an algorithm for image annotation that takes into account both image and tag features. We analyze over 5 million photos with over 20,000 visual tags. The statistics from this collection leads to good results for image tagging, relationship estimation, and generalizing to unseen tags. This is a first step in analyzing picture tags and everyday semantic knowledge. Potential other applications include generating natural language descriptions of pictures, as well as validating and supplementing knowledge databases.

Categories and Subject Descriptors: H.2.8 [**Database applications**] Data Mining
Keywords knowledge graph; social media; folksonomy

1. INTRODUCTION

An image is worth a thousand words, but which thousand? A particularly interesting aspect of this question is how people use thousands of everyday words to describe photos and videos on the web. The sheer volume of visual data presents both a challenge and an opportunity. The challenge is in understanding collective photo tagging behavior. With users posting everything from family and social events, daily life, and amateur photography, has photo tags evolved into a language of its own? Do words in natural language and words used in photo tags differ (and how, if they do)? The opportunity that comes with this challenge is to leverage such understandings and design better systems to organize and search for pictures. Specifically, insights on visual tag vocabulary will help a number of applications: better visual search, improved picture annotation and tag suggestion, and enriching encyclopedic knowledge-bases with more visual content.

Online social tagging and automated photo annotation has been a very active research area for more than a decade, yet there are three gaps in the current approaches. The first gap is between the understanding of general tagging behavior[3, 9] and the practice of automated visual tagging [7]. This is in part because tagging behavior studies rely on semi-manual approach in classifying tags and surveying tagging practices, and there has not been sufficient data to connect the findings on tag types and tagging motivations to large amounts of visual data. The second gap is between visually recognizing a pre-defined list of words, and the real-world knowledge that express relationships among these words. Visual recognition systems has mostly relied on a particular relationship, such as co-ocurrence statistics [11, 36], pre-defined lexical hierarchical structure [13], data-driven graph structures [30]. With maturing resources to encode human knowledge and everyday relationships become available [5, 21], it is conceivable that this gap will be closing. The third gap is in addressing the long tail in visual recognition. The long-tail phenomenon, that there are a large number of items not occurring in the most popular part of the distribution, has been observed picture tags [9, 40] as well as many other problem domains such as recommender systems [24]. Visual recognition efforts tend to concentrate on the head of the distribution, acknowledging that recognizing rare tags is difficult. With reliable prior information about how tags relate to each other and models that effectively shares information across different tags, this gap can also be narrowed.

This work brings together several large-scale visual and semantic resources to analyze picture tags. We propose novel methods to connect the following resources: an annotated visual ontology ImageNet, online social tagging collections from Flickr, and the commonsense reasoning engine ConceptNet. Specifically, we jointly analyze ImageNet and Flickr photos to quantify how *visually relevant* a tag is, addressing the first gap. We then link photo tags with everyday knowledge from ConceptNet. Specifically, we propose Inverse Concept Rank (ICR), a novel network inference method from co-ocurrence statistics, for estimating latent relations between visual tags and constructing better tag space similarity, addressing the second gap. Lastly, we

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502113.

propose a scalable approach for simultaneously learning and recommending many tags to photos. This approach uses the statistics from our analysis of visual and knowledge sources, and optimizes the MatchBox model, originally designed for large-scale recommender systems for long-tailed content, addressing the third gap. We report picture tag statistics on a collection of over 5 million photos, more than 20,000 words and 450K commonsense relations. Our tag recommendation approach is tested on an annotated Flickr dataset with more than 200K photos. The pilot evaluation yields over 70% precision among top-5 tags and 0.35 average precision scoring all tags. Our work analyzes data from openly available sources, and we will also make our data available at a companion website [2]. We see our method as taking the first steps towards closing three prominent gaps in social photo tags. Our method and results can lead to inquiries in several additional areas: to use tag relation for constructing sentences, to rank tags for their visual informativeness, and additional models to encode tag prior.

The main contributions of this work include:

- Novel joint analysis of three visual and semantic resources, connecting visual semantics to large-scale photo tagging behavior (Sec. 2).
- Inverse Concept Rank, a new model for estimating tag relationships from their co-occurrence statistics (Sec. 3).
- An approach for photo tag recommendation using both visual features, tag statistics, and semantic relationship (Sec. 4).
- Quantifying the visual relevance of both individual tags and tag relationships, mapping over 20,000 everyday words and their relations for the first time (Sec. 5).
- An evaluation on more than 200K photos, with promising results for photo annotation, generalizing to unseen tags, as well as relationship graph estimation (Sec. 6).

2. VISUAL AND SEMANTIC RESOURCES

We establish novel connection among three well-known visual semantic resources. This crucial connection enables us to analyze tens of thousands of photo tags and their visual and semantic relations for the first time.

The first resource is **Flickr photos and tags**. Being one of the most popular online photo sharing platforms, Flickr hosts billions of photos and makes their metadata available via its application programming interface (API). A fraction of photos are assigned one or more *tags*, a free-text string used (typically by the owner) to describe and organize the photo collection. This work also uses *owner* and photo content information on a relevant subset of Flickr photos.

The second resource is **ImageNet** [13][1], a research resource that annotates millions of pictures of nouns in the lexical database Wordnet [17]. ImageNet is organized along *synsets*, each synset is associated with a number of images, each image is deemed visually relevant to the corresponding meaning in Wordnet. There are about 14 Million annotated images in 21,841 synsets in total. These annotations were collected by crowd-sourcing, and verified by humans.

Table 1: Summary of notations.

Notation	Meaning	Defined
$\mathbf{X}, \mathbf{x}_i$	Photo(s) and their features	Sec 4
$\mathbf{t}, t_i, \mathcal{T}$	tag/concept, tag vocabulary	Sec 2
$\mathbf{s}, \mathcal{S}$	ImageNet synset	Sec 2
ξ_{t_j}	Tag informative measure	Sec 2
B, b_{ij}	Flickr bigram count	Sec 2
$\mathbf{G}, g_{ij}$	Tag/concept relation graph	Sec 2
$\mathbf{W}, w_{ij}$	Row normalized version of G	Sec 2
$\mathbf{Z}, \mathbf{z}_i$	ICR stationary distribution	Sec 2
$\mathbf{e}, \mathbf{e}_i$	Vector constants	Sec 2
ν	ICR restart distribution	Sec 2
α	ICR teleportation probability	Sec 2
$\mathbf{R}, r_{ij}$	Labels, photo-tag association	Sec 4
$\mathbf{Y}, \mathbf{y}_j$	MatchBox: tag features	Sec 4
$\mathbf{U}, \mathbf{V}$	MatchBox: latent factors	Sec 4
λ	MatchBox: regularization weight	Sec 4
κ	MatchBox: # latent dimensions	Sec 4
m, n, p, q	MatchBox: data dimensions	Sec 4
$\mathcal{L}, \mathcal{U}$	MatchBox: labeled/unlabeled sets	Sec 4
J_R, J_M	Objective functions	Sec3&4
ε	Error term	Sec3&4
$h(\varepsilon)$	Loss function	Sec3&4
ρ, θ	Tag-pair polar coordinates	Sec 5
i, j, k, u, v	General index	–

The third resource, **ConceptNet** [21], is a semantic network based on Open Mind Common Sense[2]. Statements about everyday facts are collected online via a crowd-sourcing website, and are verified using a voting system. The statements are stored in a tuple format, with each tuple consist of a pair of concepts and a relation, such as a *zipper* is *Usedfor* a *jacket*. Note that these relationships are distinct from the ontological relationship encoded by WordNet. We choose ConceptNet over other large scale knowledge bases [5, 42] as such extensive coverage of everyday terms is not found in other language-focused sources such as Wikipedia. Here, ConceptNet is used to understand the relations between everyday words that are used to describe pictures.

We acquire and process data for each of these resource in three steps. Detailed data statistics are described in Section 5. (1) Obtain the original image URLs from ImageNet, keep images that are from Flickr. Only for these images we have both a visually describable concept (from ImageNet) together with rich metadata and user-supplied tags (from Flickr). We then download the photo and their metadata from Flickr, removing photo entries that are no longer available or do not have tags. We refer to this data as the ImageNet/Flickr collection. (2) Acquire ConceptNet, both version 4[3] and version 5[4]. We keep statements that are in English and has one of fourteen most common relations, i.e.*IsA, HasA, ConceptuallyRelatedTo, UsedFor, AtLocation, DefinedAs, InstanceOf, PartOf, HasProperty, CapableOf, SymbolOf, LocatedNear, ReceivesAction, MadeOf.* The other relations are either not directly recognizable in images (e.g. *HasSubevent*) or very rare (e.g. *HasPainCharacter*). The resulting data are in list-of-triplet format, e.g. *CapableOf,butterfly,fly.* (3) Normalize the different vocabularies from the three resources. We use the “2+2lemma” dictionary [4] that includes 80,431 words. This dictionary also includes a mapping to lemmatize each word to one of

[1] http://image-net.org

[2] http://openmind.media.mit.edu

[3] http://csc.media.mit.edu/docs

[4] http://conceptnet5.media.mit.edu

32,606 "headwords". For example, *funnier, funnily, funniest* all map to *funny*.

We compute several counting statistics from these datasets to help characterize the relationships between photo tags of Flickr, visual synsets from ImageNet, and semantic relations from ConceptNet. Statistics on tens of thousands tags allow us to ask and start answering a few questions about the properties of photo tags and their relationships. Such as: If we were to use a thousand tags to distinguish visual content, which thousand should we use? Which tags are not about the picture content, but about the context that they were taken? Which tag-tag relationships are reflected in images, which aren't? Are there new relationships that we can discover from tag statistics, what would they be?

2.1 Tag statistics

The first statistic is called *visual informativeness*, derived from counts $\#(s_i, t_j)$ for tag t_j appearing in visual synset s_i. Specifically, this is counted as the number of distinct users that has issued tag t_j for any photo in s_i. Unique user count is an effective way of denoting photo tags when a user is uploading and tagging in batches [23, 28] – we found one synset contain 50 photos from the same user tagged with just *photo* and *canon*, for example. We estimate $p(s_i|t_j)$ by normalizing along each tag, i.e. $\#(s_i, t_j)/\sum_i \#(s_i, t_j)$. This visual word and tag association matrix is used to measure the visual distinctiveness of tags. For example, tags such as *pineapple* is distinctive and should only appear in a few synsets, i.e., $p(\mathbf{s}|pineapple)$ peak in synset related to *pineapple*; while *photo* and *canon* (camera brand) will be associated with many synsets, i.e. $p(\mathbf{s}|canon)$ would be flat. Conditioning on tag t_j changes the prior distribution $p(\mathbf{s})$ over visual words: the more the change, the more visually relevant t_j is. We call this quantity *visual informativeness*, and measure it with the KL-divergence [25] between the conditional probability of synsets given a tag $p(s|t_j)$ and their prior probability $p(s)$.

$$\xi_{t_j} = KL[p(\mathbf{s})||p(\mathbf{s}|t_j)] = \sum_i p(s_i)log\left[\frac{p(s_i)}{p(s_i|t_j)}\right] \quad (1)$$

2.2 Association between a pair of tags

We examine two different statistics of a pair of tags t_i and t_j. We first obtain the tag pair co-occurrence matrix $\mathbf{B}$, where count b_{ij} denotes the number of times that t_i and t_j are used to describe the same photo together in the ImageNet/Flickr collection. One natural question that arises is, does high/low co-occurrence in b_{ij} agree with high/low associations in ConceptNet?

To answer this question, we first need a measure of association for a part of concepts in ConceptNet. Denote a non-negative concept relationship graph $\mathbf{G} \in \mathcal{R}_+^{n\times n}$, where $g_{ij} > 0$ if concept i and j are related. In this paper, we obtain g_{ij} by counting the number of relations that exist between tag t_i and t_j.

We use a random walk model to describe the process of coming up with a series of tags for a picture, as this was shown to be close to human cognition in a seminal 2007 study by Griffth, Steyvers and Firl [20].

Starting from tag t_i, a user makes a transition to tag t_j (i.e. use as the next tag) with a probability proportional to g_{ij}; and return to the initial tag t_{i_1} with a fixed probability $1 - \alpha$ (i.e., finishing the tagging process for the current photo). Such a random walk process has been described as personalized PageRank (PPR) [26] or random walk with restart (RWR) [44] in the context of web search and graph mining. We note that the co-ocurrence of everyday image tags is not only driven by their conceptual relationship, but also driven by the physical colocation and other external factors. It is a simplifying assumption to capture these different types of relationships with one relationship graph, and initialize such a graph with ConceptNet and ImageNet. However, the different relations within ConceptNet, such as *Usedfor*, *LocatedNear* does account for a few different modes of co-occurences.

Given a relation graph $\mathbf{G}$, we generate a stochastic version $\mathbf{W}$ of the graph by normalizing along each row.

$$w_{ij} = \frac{g_{ij}}{\sum_k g_{ik}} \quad (2)$$

The RWR transition probability $\hat{\mathbf{W}}$ is computed from $\mathbf{W}$ as as a weighted combination between taking a step according to $\mathbf{W}$ and jumping back to initial distribution ν, with a *teleportation* constant $\alpha \in [0, 1]$, and $\mathbf{e}$ as an all-1 vector.

$$\hat{\mathbf{W}} = \alpha\mathbf{W} + (1 - \alpha)\mathbf{e}\nu^T \quad (3)$$

Denote vector z as the stationary probability distribution of Markov chain $\hat{\mathbf{W}}$, i.e, the personalized page rank vector. z satisfies:

$$\begin{aligned} \mathbf{z} = \hat{\mathbf{W}}^T\mathbf{z} &= [\alpha\mathbf{W}^T + (1-\alpha)\nu\mathbf{e}^T]\mathbf{z} \\ &= \alpha\mathbf{W}^T\mathbf{z} + (1-\alpha)\nu \end{aligned} \quad (4)$$

Note that the last step follows from normalization $\mathbf{e}^T\mathbf{z} = 1$. The initial distribution ν_i for starting from the note i is set as a vector with a 1 in the i^{th} position and 0 elsewhere. Varying the starting node i will give a series of different stationary distribution z_i, forming a matrix $\mathbf{Z} = [\mathbf{z}_1^T; \ldots; \mathbf{z}_i^T; \ldots]$. We use $\mathbf{z}_{ij}$, the stationary probability of being at node j if started from node i, as the association between two concepts i and j under ConceptNet.

Detailed observations about ξ_t, as well as comparisons between $\mathbf{B}$ and $\mathbf{Z}$ are presented in Section 5, and learning algorithms that are motivated by such observations are in Sections 3 and 4.

3. INVERSE CONCEPT RANK

It would be desirable if concept-concept, or tag-tag relationships could explain the process for generating photo tags. As seen in Figure 3 and Section 5.2, however, that there are a number of mismatched cases between tag occurrence and concept relationships, this is in part due to incomplete relations in ConceptNet, noisy counts from both visual and non-visual tags, non-visual relations that do not manifest in photos, and person- and event- specific factors that are not explained by everyday semantic knowledge. In this section we propose a model to address the first two factors, by estimating latent relations, and unsmooth probability observations for better concept similarity.

3.1 An inverse random walk model

Section 2.2 described a random walk process of generating photo tags, and the resulting tag distribution can be computed by personalized PageRank. Here we are facing an inverse problem: if the stationary distributions of such a random walk is observed (via bigram statistics $\mathbf{B}$ and its normalization $\bar{\mathbf{Z}}$), but the underlying graph $\mathbf{G}$ is hidden or only partially known. The goal is to "recover" a graph $\mathbf{G}$

that generates $\bar{\mathbf{Z}}$. Hence we name this model Inverse Concept Rank (ICR).

Let observed stationary distribution be $\bar{\mathbf{Z}} = [\bar{\mathbf{z}}_1^T; \ldots; \bar{\mathbf{z}}_n^T]$, where each $\bar{\mathbf{z}}_i$ is the personalized page rank vector starting from node i. The goal of the ICR model is to find an optimal relationship graph $\mathbf{G}$, from which an RWR process will generate $\mathbf{Z}$ that is close to $\bar{\mathbf{Z}}$. The objective function is define as follows:

$$\begin{aligned} \min_{\mathbf{G}} J_R &= \frac{\alpha_R}{2} tr(\mathbf{G}^T\mathbf{G}) + \sum_{i,j} h(|z_{ij} - \bar{z}_{ij}|) \\ & \quad RWR(\mathbf{G};\alpha) \rightarrow \mathbf{Z} \\ s.t. & \quad \mathbf{G} \geq \mathbf{G}_0 \end{aligned} \tag{5}$$

Here $RWR(\mathbf{G};\alpha)$ is the RWR process on $\mathbf{G}$ with parameter α, as in Equations 2–4; $|z_{ij} - \bar{z}_{ij}|$ is an error term to be minimized; h is a monotonic loss function – common choices include L2, hinge loss, etc.; $\alpha_R tr(\mathbf{G}^T\mathbf{G})$ is a regularizer of the Frobenius norm of $\mathbf{G}$ weighted by hyper parameter α_R. This regularizer favors smaller graph weights. $\mathbf{G}_0$ is an initial graph with non-negative edge weights, the inequality constraints are element-wise on $\mathbf{G}$.

3.2 Optimizing Inverse Concept Rank

The ICR objective (5) has no closed-form solution, it is non-linear and non-convex with respect to graph entries g_{ij} due to the RWR process. Numeric solution such as quasi-Newton method can be used to minimize this objective. In particular, we use large-scale non-linear solver L-BFGS-B [32] with closed-form gradient computed as follows.

The partial derivative of J_R with respect to g_{ij}, $(i,j) \in \{1,\ldots,n\}^2$ can be expressed in terms of z_{uv}, entries in the stationary distribution matrix and their associated error terms $\varepsilon_{uv} = z_{uv} - \bar{z}_{uv}$, with $(u,v) \in \{1,\ldots,n\}^2$. Applying chain rule and matrix identities [35] to Equation (5) gives:

$$\frac{\partial J_R}{\partial g_{ij}} = \alpha_R g_{ij} + \sum_{u,v} \frac{\partial h(\varepsilon_{uv})}{\partial \varepsilon_{uv}} \sum_k \frac{\partial z_{uv}}{\partial w_{ik}} \frac{\partial w_{ik}}{\partial g_{ij}} \tag{6}$$

Each of the three terms in Equation (6) can be computed in closed form. $\frac{\partial h(\varepsilon)}{\partial \varepsilon}$ is easily computed for all common loss functions. For L2 loss $\frac{\partial h(\varepsilon)}{\partial \varepsilon} = 2\varepsilon$, for hinge loss we can use a sub-gradient, or a smoothed quadratic surrogate. $\frac{\partial w_{ik}}{\partial g_{ij}}$ is computed from the normalization equation (2):

$$\frac{\partial w_{ik}}{\partial g_{ij}} = \frac{\delta(k=j)}{\sum_l g_{il}} - \frac{g_{ik}}{(\sum_l g_{il})^2}$$

Here $\delta(k=j)$ is the indicator function, and the first term only appears when $k=j$.

The remaining term $\frac{\partial z_{uv}}{\partial w_{ik}}$ is the sensitivity of stationary distribution $\mathbf{z}$ with respect to the Markov matrix $\mathbf{W}$. Golub and Meyer [18] showed that for a irreducible Markov chain,

$$\frac{\partial \mathbf{z}}{\partial w_{ij}} = \alpha \mathbf{z} \mathbf{e}_i \mathbf{e}_j^T (\mathbf{I} - \hat{\mathbf{W}})^\sharp. \tag{7}$$

Here $\mathbf{A}^\sharp$ is the group inverse of $\mathbf{A}$, a unique matrix satisfying $\mathbf{A}\mathbf{A}^\sharp\mathbf{A} = \mathbf{A}$, $\mathbf{A}^\sharp\mathbf{A}\mathbf{A}^\sharp = \mathbf{A}^\sharp$, and $\mathbf{A}^\sharp\mathbf{A} = \mathbf{A}\mathbf{A}^\sharp$. $\mathbf{e}_i$ is a vector with 1 in the i-th position and the rest being zeros. The group inverse can be computed with QR decomposition. Let $(\mathbf{I} - \hat{\mathbf{W}}) = \hat{\mathbf{Q}}\hat{\mathbf{R}}$ be a QR factorization, $\hat{\mathbf{R}}$ must have the form

$$\mathbf{R} = \begin{bmatrix} \hat{\mathbf{U}} & -\hat{\mathbf{U}}\mathbf{e} \\ \mathbf{0} & 0 \end{bmatrix} \tag{8}$$

The group inverse is given by

$$(\mathbf{I} - \hat{\mathbf{W}})^\sharp = (\mathbf{I} - \mathbf{e}\mathbf{z}) \begin{bmatrix} \hat{\mathbf{U}}^{-1} & 0 \\ \mathbf{0} & 0 \end{bmatrix} \mathbf{Q}^T (\mathbf{I} - \mathbf{e}\mathbf{z}) \tag{9}$$

Note that we represent ConceptNet as a undirected graph, this necessarily make $\mathbf{G}$ irreducible, and satisfies the condition of Equation 7. Note that the computational complexity of each gradient step is cubic in graph size n, with QR decomposition taking $O(n^3)$ (but only need to be done once), and the summation in Equation 6 also taking $O(n^3)$. Quasi-Newton method finds a local minima in the J_R that is close to starting point $\mathbf{G}_0$. In practice, we run ICR within each synset, saving computational time and also retaining synset-specific relations.

Intuitively, graph $\mathbf{G}$ is parsimonious and will be free from noisy chain-cooccurrences observed in $\mathbf{B}$. For example, *hammerhead (shark), atlanta* is among the top tag-pairs in synset *n01494475*, but this is because *"shark isLocatedAt aquarium"*, and *"an aquarium isLocatedAt Atlanta"*, not because *"hammerhead LivesIn Atlanta"* – impossible for an inland city. Recovering the underlying graph $\mathbf{G}$ given observations $\bar{\mathbf{Z}}$ can have a number of applications. A direct use would be to interpret $\mathbf{G}$, such as finding and filling in missing entries in ConceptNet. Another type of application is to use $\mathbf{G}$ as an *un-smoothed* version of tag correlation, this can serve as feature for automatic or semi-automatic picture tagging. Observations and evaluations of the ICR model on Flickr data will be presented in Sections 5 and 6.

4. TAG RECOMMENDATION

One application of tag statistics (in Sections 2.1 and 5.1) is to serve as a similarity metric between different visual concepts. Tag similarity is a prior information that can help infer unknown picture tags that has too few or no training data, and to regularize individual tag predictors. We draw an analogy between photo tagging and collaborative recommendation problems that simultaneously infer the preferences to a large number of different items. Latent space models are effective choices for this purpose [24].

We denote each image as a feature vector $\mathbf{x}_i \in \mathbb{R}^n$, and each tag as t_j, $i = 1,\ldots,n$, $j = 1,\ldots,m$. Denote a tag matrix as $\mathbf{R} \in \mathbb{R}^{n \times m}$, each element $r_{ij} \in \{1,-1,0\}$ represent the label of a photo-tag pair $(\mathbf{x}_i, t_j)$. Here 1 means t_j describes $\mathbf{x}_i$, -1 means t_j does not describe $\mathbf{x}_i$, and 0 means unknown. Predicting unknown entries in R is akin to a recommendation problem [24], and one effective approach that takes into account item and tag features is the MatchBox model [41]. It approximates R with two latent factors: the photo factor $\mathbf{U} \in \mathbb{R}^{\kappa \times p}$ on the p-dimensional photo features and the tag factor $\mathbf{V} \in \mathbb{R}^{\kappa \times q}$ on the q-dimensional tag features. Denote the set of labeled photo-tag pairs as $\mathcal{L} = \{(i,j), \text{where } r_{ij} = \pm 1\}$ and unlabeled pairs are $\mathcal{U} = \{(i,j), \text{where } r_{ij} = 0\}$. This problem is expressed as a minimization problem for mean-square loss in R, regularized by the norms of latent factors $\mathbf{U}$ and $\mathbf{V}$,

$$J_M = \frac{1}{2} \sum_{(i,j) \in \mathcal{L}} (r_{ij} - \mathbf{x}_i^T \mathbf{U}^T \mathbf{V} \mathbf{y}_j)^2 + \frac{\lambda}{2}(tr(\mathbf{U}^T\mathbf{U})) + tr(\mathbf{V}^T\mathbf{V})) \tag{10}$$

Here $\mathbf{x}$ and $\mathbf{y}$ are features describing pictures and tags respectively, and λ and number of latent dimension κ are hyper-parameters. Note that objective function J is non-convex in $\mathbf{U}$ and $\mathbf{V}$, but is convex (quadratic) in either $\mathbf{U}$

or $\mathbf{V}$ if we hold the other fixed. We adopt an alternating gradient descent [24] approach to find a local minima in J. In short, we take derivatives of $\mathbf{U}$ and $\mathbf{V}$ in turn while holding the other constant. Then we apply gradient descent in a round-robin fashion until a local minima is reached for all parameters. This is implemented with an L-BFGS-B solver [32] with gradients defined as follows [35]:

$$\frac{\partial J_M}{\partial \mathbf{U}} = -\sum_{(i,j)\in\mathcal{L}} \varepsilon_{ij}\mathbf{V}\mathbf{y}_j\mathbf{x}_i^T + \lambda\mathbf{U} \qquad (11)$$

$$\frac{\partial J_M}{\partial \mathbf{V}} = -\sum_{(i,j)\in\mathcal{L}} \varepsilon_{ij}\mathbf{U}\mathbf{x}_i\mathbf{y}_j^T + \lambda\mathbf{V}$$

Note that $\varepsilon_{ij} = r_{ij} - \mathbf{x}_i^T\mathbf{U}^T\mathbf{V}\mathbf{y}_j$ is a shorthand for the current error to predict r_{ij}. Also note that the computational cost of the objective function and its derivatives is linear in the number of labeled photo-tag pairs $|\mathcal{L}|$, linear in the number of labeled images n and tags m, linear in feature dimensions p and q, and quadratic in the (typically small) latent dimension κ, i.e. $O(|\mathcal{L}|mnpq\kappa^2)$. Compared to Wasabie [49], our bilinear model incorporates additional tag features. We currently use a square loss, on ε_{ij}, we observed that hinge loss performs comparably, and similarly, a ranking loss can also be incorporated.

4.1 Photo Matchbox applications

This latent-space model of photos and tags has many applications, this paper evaluates three: (A) "Collaborative" tagging. Minimizing equation (10), and then use $\hat{R} = \mathbf{x}^T\mathbf{U}^T\mathbf{V}\mathbf{y}$ to predict entries for the unlabeled set $\mathcal{U}$. (B) Multi-tag photo labeling. This applies the Matchbox model to a new photo x with no existing tags. The best labels $R_{\mathbf{x}}$ for x is obtained by directly applying the parameters without needing to re-optimize for U and V, i.e. $\hat{R}_{\mathbf{x}} = \mathbf{x}^T\mathbf{U}^T\mathbf{V}\mathbf{Y}$ with the label estimate $\hat{R}_x$ being a multiple linear combinations on feature x. (C) Inferring unknown tags. This is the dual of problem (B) above, where $\hat{R}_y = \mathbf{x}^T\mathbf{U}^T\mathbf{V}\mathbf{y}$ for an unknown tag with feature $\mathbf{y}$. Evaluations of these applications can be found in Section 6.2.

Note that there are different choices in supplying the photo and tag feature $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{X}$ can include low-level perceptual features [12], bag-of-local descriptors [45], or mid-level representations such as ObjectBank [27]. $\mathbf{Y}$ can include any feature that describe a tag, in this paper, we consider the synset-tag association matrix $\mathbf{T}$, top m rows sorted by *visual informativeness* ξ. We also consider graph structures from *un-smoothing* co-occurrence observations, as described in Section 3.

5. DATASET AND OBSERVATIONS

We acquire and process ImageNet and ConceptNet datasets as described in Section 2. Out of 14.2 million photos on ImageNet, 6.3 million are from Flickr, and 5,107,147 photos (or $\sim$ 35%) have one or more tags that can be found in the dictionary, as of March 2012. Out of all 21K+ synsets, 13,288 contain more than 5 tagged Flickr photos, and thus kept for analysis. We map both Flickr tags and ConceptNet terms using the *2+2lemma* dictionary [4]. We disregard the multi-word terms and any terms that did not match. We also prune words that appeared less than 5 times in the entire ImageNet/Flickr collection. This has left us a vocabulary of 20,366 words from Flickr tags, of which 7,796 also appears in ConceptNet.

5.1 Observation on tags

We compute the *visual informativeness* measure ξ (Eq. 1) from the co-ocurrence matrix of 13,288 synsets × 20,366 tags. These tags are visualized with its frequency $|t|$ vs *visual informativeness* ξ_t.

The middle overview graphs of Fig. 1 and the yellow background in Fig. 2 are scatter plots of all 20K+ tags, we can see that there is a wide range of *visual informativeness* value for tags of similar frequency. Fig. 1 (A) and (B) zoom in on the overall scatter plot, as indicated by the overview icons, showing the lower and upper envelope of the tag scatter, respectively. For Fig. 1, in particular, the light blue tags contain an annotated (partial) list of 194 food names (A), and 130 place names (B). These food and place names are collected from the top tags across a few dozen food- and travel-related Flickr groups, and then manually annotated by one of the authors. Overall, the upper and lower regions of the tagger scatter shows a clear trend of concrete, visual nouns (2A, 1B) versus abstract and non-visual tags (2B, 1A). This method for correlating a visual resource (ImageNet) with generic tagging resource (Flickr) can be used to re-rank tags in an image, or to decide which 1000 visual classifiers to train (first) for better usability and performance.

5.2 Observation on tag pairs

We obtain bigram statistics $\mathbf{B}$ and RWR probability $\mathbf{Z}$ on ConceptNet graph, as described in Section 2.2. There are 1,507,457 tag pairs that appeared in at least 5 images over the ImageNet/Flickr collection, a scatter plot of b versus z is in Fig 3 (left). There are three salient regions in this plot, quantified by transforming the $b - z$ plane to normalized polar coordinates. We first normalize the two quantities to be between 0 and 1: we take the log on bigram counts, and then divided by the log of the maximum bigram to obtain $\tilde{b}$; we divide z by the maximum of all z to obtain $\tilde{z}$. The polar coordinate of a point (b, z) is:

$$\rho = \sqrt{\tilde{b}^2 + \tilde{z}^2},\ \theta = arctan(\frac{\tilde{z}}{\tilde{b}}).$$

Intuitively, we notice three salient regions according to different values of θ, marked as I, II, and III on Fig 3 (left). Top tag-pairs of each region are shown on Fig 3 (right), obtained as the top 20 pairs by descending values of ρ, within the angles ranges specified according to θ. Region I ($\theta \sim 45^\circ \pm 7.5^\circ$) can be called *concurrence*, when we tend to observe high tag co-ocurrence with high associations in ConceptNet. Examples in Fig 3 (right) shows intuitively related tags that are also used to describe objects and scenes in photos, such as *sea* and *water*, *book* and *library*. Region II ($\theta \in [0^\circ \pm 15^\circ]$) can be called *new visual relations*, these tag-pairs frequently co-occur in photos but are not strongly associated in ConceptNet. The examples show intuitively related tags that are also used to describe pictures, such as *animal* and *nature*, *flower* and *spring*. Observations in these region can potentially be used to generate new statements for semantic networks including ConceptNet. Region III ($\theta \geq 68^\circ$, the maximum θ is 72°) can be called *non-visual relations*, these tags are strongly associated in ConceptNet statements, but do not frequently appear together in photos. Examples show abstract words (e.g., *duty*, *domicile*) and as events *rent* and *house* that are difficult to depict in a single image.

The joint examination of commonsense and photo tag co-occurrence provides a number of interesting observations.

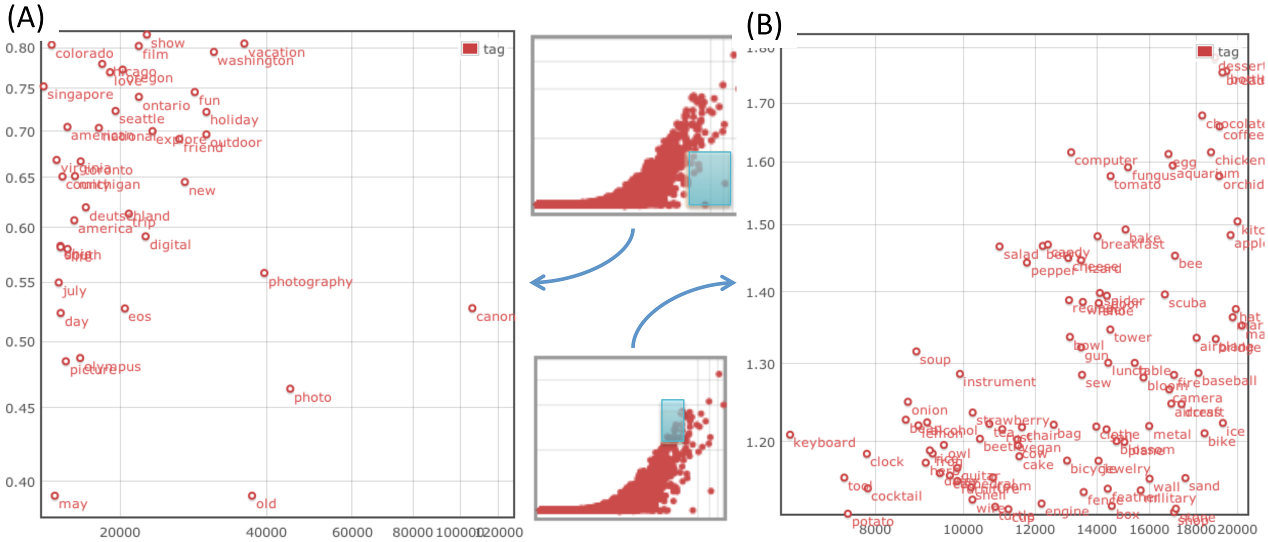


Figure 1: Zoomed-in view of the tag scatter in Fig 2. (A) An area with non-informative tags, containing abstract nouns, adjectives and place names; (B) An area with informative tags, mostly various object names. (For better readability, please view in color and with magnification).

They can be starting points for composing a sentence to describe photos, by providing a basis to include abstract and non-visual tags with visual ones. On the other hand, this analysis also exposes statistical biases in the data. Flickr bigrams are certainly biased by the sheer frequency of a tag, e.g. *blue* appeared many times in Figure 3, but other colors did not. The number of ConceptNet4 tuples (450K) is about an order of magnitude less than the Flickr bigrams (>5M total, 1.5M with counts >5). The example in regions II shows the prospect of better tag association from data. This is one of our motivations for the ICR model – leveraging Flickr as a source to infer the missing relationships in the ConceptNet graph.

6. EXPERIMENTS

We first evaluate the ICR algorithm alone for inferring concept relation graph, and then evaluate image tagging with a number of different tasks and tag features.

For ConceptNet, we keep two versions of the relation graph with 7,796 nodes: ConceptNet4 (CN4) has 450K total relations and 51,576 relations after tag and relation filtering (Sec 2); ConceptNet5 (CN5), released in 2012, has 69,120 relations on the same set of tags (or 1/3 new relations, "CN5-new" for short).

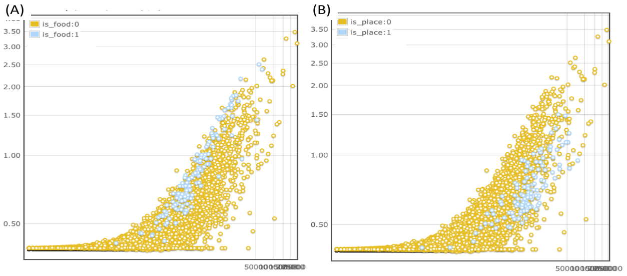
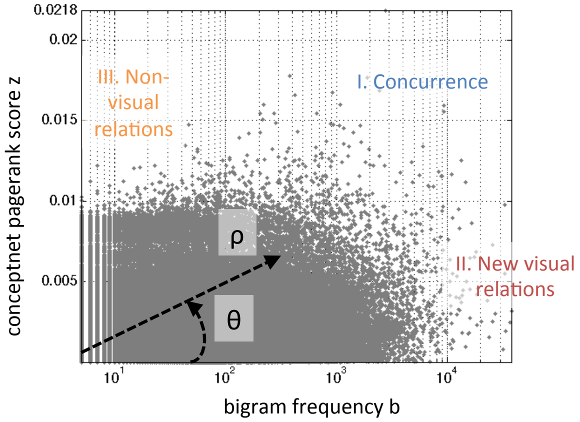


Figure 2: Scatter plot of log-frequency vs visual informativeness ξ for the ImageNet-Flickr collection. (A) Common food names against all tags; (B) Common place names against all tags;

We measure image tagging performance on the NUS-WIDE [43] dataset, containing 269K images collected from Flickr, annotated with 81 visual tags, split 60-40 into training/testing sets. This dataset is suitable for our evaluation, not only as it is from the same source as the ImageNet-Flickr collection, More importantly, it contains almost complete yes/no judgment including negatives, most other datasets including ImageNet only has positive examples, and provide no negative judgement on picture-tag tuple. We download photos directly from Flickr with the provided URL, and then extract an 177-dimensional ObjectBank feature from each image. ObjectBank [27] produces for each object a response map containing 2 views x 6 scales x 3 pyramid levels $(1+2^2+4^4 = 21)$, and the responses for each object are aggregated by taking the maximum. This vector is used to populate X in MatchBox. We map the visual tags annotated by NUS-WIDE [12] to ImageNet synsets using dictionary lookup on all words in the synset. We removed a few that either has no matches (e.g. *sunset*) or matched synset does not have more than 50 photos. We also filtered a few tags that are too generic for ImageNet (e.g. people, the top of a wordnet tree with 2800+ synsets). We are left with 63 visual tags as targets for MatchBox annotation.

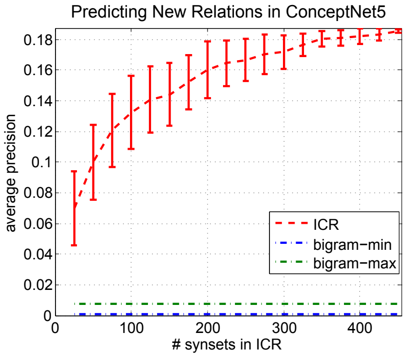
6.1 Inverse Concept Rank Evaluation

We use the ICR algorithm to learn the underlying concept graph, starting from CN4 and using bigram counts as observations $\bar{\mathbf{Z}}$. The evaluation target are relations in CN5-new, we note that these relations are incomplete (consistent with partial labeling in typical IR evaluations), but they are unknown to CN4 prior to model learning. We restrict ICR to work within each synset, since this preserves the particular relation context, and makes the algorithm execute fast – we note that ICR on different synsets can be trivially pararellized, and each synset typically take a few minutes on one CPU for ∼100 related tags from 100s to



Region I		Region II		Region III	
ocean	water	animal	nature	toil	work
sea	water	flower	nature	house	rent
build	city	bird	nature	build	thing
tree	wood	nature	wildlife	employ	work
house	room	bird	wildlife	basement	build
office	work	flower	yellow	salary	work
drink	water	animal	wildlife	build	carpet
city	house	black	white	duty	work
build	construction	flower	pink	domicile	house
eat	food	green	red	pay	work
city	town	green	yellow	build	pen
build	home	flower	white	album	work
house	window	girl	woman	work	write
job	work	blue	red	build	you
door	house	flower	spring	act	work
build	office	garden	nature	construct	house
build	structure	flower	red	sing	work
fun	play	green	white	build	supermarket
book	library	flower	purple	alcohol	build
game	play	bug	insect	here	place

Figure 3: (Left) A scatter plot of Flickr bigram frequency b and ConceptNet pagerank z with three salient angular regions. (Right) Top 20 tag-pairs of each region by ρ. See description in Sec 5.2, best viewed in color.



W1	Top10 –W2									
container	winery	alcohol	window	egg	head	lamp	table	oil	reflection	red
cross	animal	temple	chapel	pray	decoration	faith	band	crucifixion	red	white
cut	work	meat	tree	hole	rip	small	surgery	band	red	reflection
town	landmark	friend	rural	zoo	desert	field	bright	god	downtown	marina
army	national	weapon	jeep	infantry	big	glory	canon	pistol	museum	black
business	man	travel	suit	hotel	small	retail	reflection	portrait	town	design
corn	horse	cup	meal	fruit	carrot	soup	home	red	rural	holiday
guard	military	fence	weapon	sentry	work	canon	black	police	orange	brick
party	friend	night	play	fiesta	car	black	red	white	age	blue
rest	vacation	new	home	smile	body	perch	park	black	travel	nature

Relations in both ICR-400 and ICR-50 Relations in ICR-400 and not in ICR-50 Relations not in Concept5

Figure 4: Result of ICR on ILSVRC synsets. (Left) AP of new CN5 relations with varying number of input synsets. (Right) Example top relations from ICR. See descriptions in Sec 6.1, best viewed in color.

1000s photos. We use 1000 synsets from ILSVRC [1] in this evaluation, of which 462 synsets have sufficient data in the ImageNet/Flickr dataset. The graph $\mathbf{G}$ from different synsets are then dimension-aligned and aggregated by summing the weights. Hyper-parameters $\alpha = 0.5$, $\alpha_R = 10$, set with cross-validation.

Fig. 4(left) plots average precision of CN5-new relations after aggregating ICR on 25,50,…, 462 synsets. The error bars are generated with 10 random permutations with which the synsets are aggregated. We can see that ICR produces successively better predictions of CN5-new, and its predictions are significantly better than the best of using bigrams (the green baseline). Fig. 4(right) contains 10 example words (W1) and their respective top 10 related words ranked by ICR with 50 and 400 synsets (ICR-50 and ICR-400 for short). We can see that ICR-400 recovers many correct relations (red) in addition to those from ICR-50 (burgundy). We note that the learned relations can be: a) in ground-truth and reasonable (*army-weapon*), b) not in ground-truth but can be potential additions to ConceptNet (*rest-park*), c) specific to photography (*container-reflection*), and d) in ground-truth but no strong immediately connection *party-car*), due to noise in ConceptNet. Among the 333 words that have at least 10 relations in CN5-new and ICR-50, 52.5% and 76.9% have precision@10 $\geq$ 0.5 on ICR-50 and ICR-400, respectively.

6.2 Picture Tagging Evaluation

We use the wordnet-tag association matrix as tag feature $\mathbf{Y}$. We rank the tag-feature dimensions by the informative measure ξ_t, and take a subset of these dimensions from the top, i.e. with a threshold on the y-axis in Fig. 2. We optimize the objective function in Eq. 10. Positive entries of r_{ij} are kept, negative entries are subsampled up to 8x number of positive samples for each tag, the rest are treated as the unknown set $\mathcal{U}$. We use cross-validation to choose regularization weight λ, the number of latent dimensions κ, and the dimensionality of tag features q. Each of these parameters are shown to be performance-insensitive within a range, we also found hinge loss and square loss to perform similarly. We report results with square loss and the best configuration $\lambda = 100$, $\kappa = 5$ and $q = 150$. Training and testing on the entire NUS-WIDE dataset takes 2$\sim$3 hours on one CPU core. Details of model parameter tuning can be found in the supplemental material [2].

Fig. 5 compares MatchBox model to a number of baseline approachches. Fig. 5(A) reports *micro-* average precision where each (image, tag) pair $\hat{r}_{ij}$ is considered as a retrieval target; (B) reports macro AP (or mAP), mean average precision over different tags; (C) measures tag recommendation performance with precision @ 5 tags, for 1000 randomly selected test images with $\geq$5 tags. KNN and SVM are the visual-only results reported for NUS-WIDE [43], mAP of SVM classifier is shown to be below 0.06 and on par with

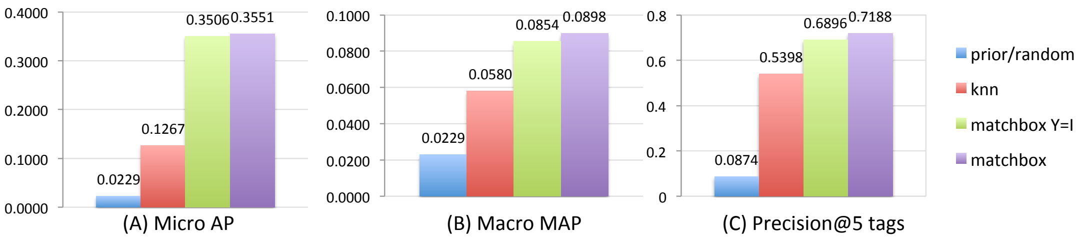


Figure 5: Image annotation results on NUS-WIDE dataset.

KNN. We can see that the bilinear matchbox model with tag features significantly outperforms the k-nearest neighbor (KNN) baseline and SVM reported earlier [43]. Matchbox with co-occurrence tag feature $\mathbf{Y}$ also outperforms the same bilinear model with identity features ($\mathbf{Y} = \mathbf{I}$) that do not share information across different tags.

6.2.1 *Tagging with unseen tags*

To evaluate the role of tag feature $\mathbf{Y}$ as a prior for tag-space similarity, we test the MatchBox model on new tags without training instances. This is similar to novel word recognition presented by Mitchell et al. [31], here we generalize using image features and tag usage, rather than using brain signals. Fig. 6 (A) reports annotation performance of the 20 most frequent tags without training data. We remove all positive instances of the tag from training set $\mathcal{L}$, and optimizes the same MatchBox objective. This model rely solely on $\mathbf{Y}$ to generalize across tags. We can see that the performance without training on the current tag are on average almost as good as those with training data, sometimes even better, such as *clouds* and *animal*. We also try to quantify the effects of tag features alone, and found that simple voting from the top-3 similar tags (T/1-voting) has comparable performance as those using feature similarities in $\mathbf{Y}$ (T/1). Fig. 6 (B) shows the top returned results for a number of free-text tags that are not in the NUS-WIDE labeled tag set. We can see that the top images of *travel* reflects two primary means of traveling: *cars* and *bicycle*; *blue* captures the color appearance of *water* and *sky*, and *architecture* is closet to learned tags *buildings, tower*, and *temple*. These results show that tag features indeed help generalize new, untrained tags in non-trivial ways.

6.2.2 *ICR for tagging*

We use the output of ICR for photo tagging. We start ICR on 93 synsets that directly map to, or mentions the NUS-WIDE tags in its name from CN5 to obtain a more complete graph estimate. We aggregate the set of resulting subgraphs $\hat{\mathbf{G}}^i$ over different synsets s_i by taking the max across elements indexed by the same tag-pair, generating $\hat{\mathbf{G}} = \{\hat{g}_{uv}\}$, where $\hat{g}_{uv} = \max_i\{\hat{g}^i_{uv}\}$. We then take elements of $\hat{\mathbf{G}}$ to the same row- and column-dimensions of the original tag feature $\mathbf{Y}$, denoted as $\hat{\mathbf{G}}_{\mathbf{Y}}$. We use the sum of $\mathbf{Y}$ and $\hat{\mathbf{G}}_{\mathbf{Y}}$ as the tag feature for MatchBox.

Figure 7 shows the mAP (macro average) for all 63 and the least frequent 43 tags. *cooc* is the original synset-tag co-occurrence feature (top bar in Fig. 5), *cn5* is using ConceptNet5 as-is for $\mathbf{G}_{\mathbf{Y}}$, and *icr5* uses $\hat{\mathbf{G}}_{\mathbf{Y}}$ after ICR. We can see that the adding *unsmoothed* tag relations from either *cn5* or *icr5* improves performance, and *icr5* further improves upon *cn5*. We also note that despite a small different mAP value, the improvement of *icr5* from *cooc* is statistically significant. Moreover, the relative improvement on the 43 rare tags (mAP-bottom43, 7.3%) is more than that across all tags (mAP-all, 3.6%). Further improvement can be expected when the target tag space is larger, and we leave this to future explorations.

7. RELATED WORK

A number of different research topics are related to our work, including understanding social tags, connecting words to pictures, picture tagging with multiple input modality and tag structure, multimedia knowledge sources, and network inference.

The recent rise of social tagging has elicited much research curiosity. Ames et al. [3] conduct surveys to understand motivation for user tagging on Flickr; Sigurbjornsson and Zwol [40] studies overall Flickr tag statistics; Bischoff et al. [9] classifies social tags into a few distinct categories, and Overellet al. [34] uses wikipedia for tag classification. A number of other work focus on the analyzing and recognizing specialized social tags, such as landmarks[23], events and places [39], or location names and proper nouns [48]. Weinberger and Slaney [48] propose a scalable approach to identify tags with ambiguous word senses, and suggest co-ocurring tags to disambiguate among its multiple meanings. All of these work provides useful insights and techniques for analyzing Flickr data, however none has profiled the use of everyday words as tags.

There are many methods for generating tags from visual and word input. Nearest-neighbor methods have been popular for large amounts of training data, photo tagging methods have relied on voting with photo metadata [40], or voting from neighbors in visual feature space and then aggregate their flickr tags [28]. Quelhas et al.[38] uses latent semantic analysis to extract visual scene patterns; Wang et al. [46] use bag-of-keypoints to illustrate a tag; Liu et al. [29], Tang et al. [43] and Qi et al. [37] design several algorithms for propagating picture tags from both visual content and noisy text. There has been competing approaches for learning multiple tags simultaneously, including full-connected graphical models [36], tree-structured graphical models derived from co-ocurrence data [11], pre-defined lexical hierarchies [14], co-ocurrence and colocations [30], Object Relation Network [10] for representing the most probable meaning of the objects and their relations in an image, and bilinear large-scale annotation model with approximate-rank loss called Wasabie [49]. A key differentiation of our work is two ways of representing prior knowledge, and a method to incorporate them as tag features for learning. Our proposed method is built upon recent developments in matrix factorization and social collaborative filtering [24, 41], while the particular application in image tagging is new.

Another related topic is to quantify the correspondence of visual elements and their descriptions. This include a "vi-

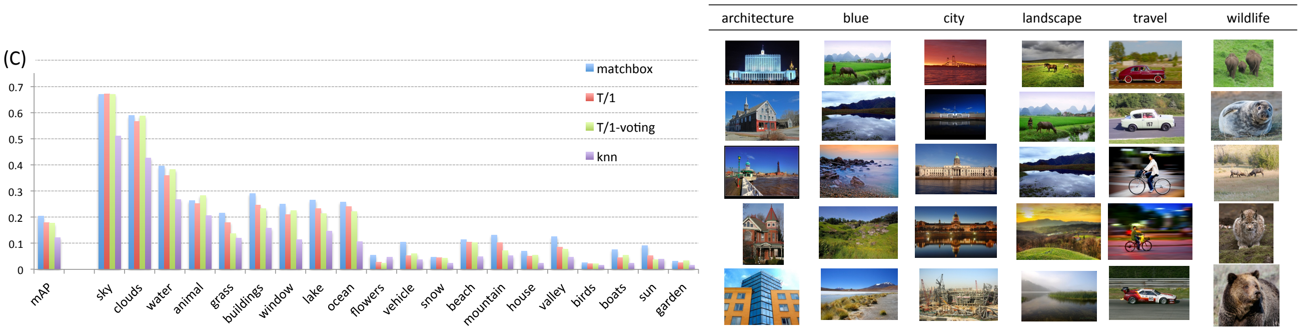


Figure 6: Annotation for unseen tags. (A) AP and mAP for the most frequent 20 NUS-WIDE tags with or without training instances (T/1), and nearest-neighbor voting in tag feature space (T/1-voting). (B) Top 5 returned images for freetext tags not in the NUS-WIDE set. See discussion in Sec 6.2.1.

sualness" measure via image region entropy [50], with both entropy and purity measures [22], determining representative web images [47], or a method to correlate visual and semantic similarities on ImageNet [15]. Recently the Stony Brook and UMD team worked on a complementary pair of problems: to predict the likelihood that a visual object is described in natural language from its properties (such as category, size and position) [8], and to detect which word(s) in a natural sentence has visual correspondence in the image that it is describing [16]. Compared to these interesting studies, our work quantifies the visual informativeness of a word by examining correlations in tagged collections alone, and do not rely on visual analysis on individual images.

Our work in mapping tag relations is inspired by research on representing real-world knowledge from web and multimedia sources. Several well-known visual ontologies have been manually designed by experts and covers a few dozen to a few hundred nodes [33]. Recently, ImageNet [13] uses images to illustrate a well-known language ontology WordNet [17], which contains tens of thousands of nodes, and has took over a decade to construct by a team of linguistic experts. Real-world knowledge databases has enlisted help from the general online crowd, YAGO and DBPedia [5, 42] are built upon Wikipedia, and ConceptNet [21] consist of crowd-sourced statements from a website over a number of years. This work can complement some of these large ontologies by providing statistical information about how words and word-to-word relations are grounded in images and used to describe images.

Inferring an underlying network from observed statistics is an interesting problem recently starting to receive more attention. The NetInf algorithm [19] estimates a directed network from observed diffusion traces, the supervised random walk [6] model learns a regressor on network parameters based on nodes reached by random walks. Our problem differs from the above as it estimates graph weights given aggregate stationary distribution statistics, not traces. This can be seen as the inverse of the well-known page-rank problem [26]. To the best of our knowledge, no solution to this problem exists.

8. CONCLUSIONS

We propose novel methods to analyze photo tags and tag relationships, using data from Flickr, ImageNet and ConceptNet. A novel network inference algorithm, ICR, is designed to estimate latent relationships from tag co-

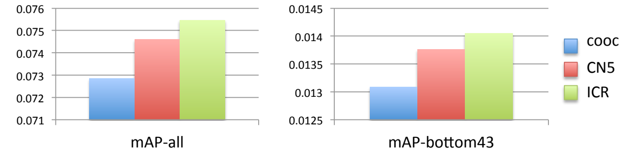


Figure 7: mAP for all and the least frequent 43 tags, with or without ConceptNet and ICR features. Note that the improvement from Co-occurence to ICR feature is statistically significant (p=0.008 from a paired t-test).

ocurrence. We obtain tag statistics on thousands of tags from millions of images. This allows us design an efficient tagging algorithm to simultaneously model many tags with image- and tag- features. The proposed tagging algorithm generalizes to unseen tags, and is further improved upon incorporating tag-relation features obtained via ICR. Core novel aspects of our work are in quantifying visual tag use and tag relations from social statistics, algorithms for network inference from aggregated occurrences, and using these insights for large-scale picture tagging.

The limitations of this work point to several directions for improvement and future work, such as: techniques to better incorporate multi-word terms and out-of-vocabulary words; advanced NLP techniques for learning word relations from free-form text; evaluation of latent concept relation suggestion, and predicting the type of relations.

Acknowledgments

This research was partially undertaken on the NCI National Facility in Canberra. NICTA is funded by the Australian Government through the Department of Broadband, Communications and the Digital Economy and the Australian Research Council.

9. REFERENCES

[1] ILSVRC, 2012. http://www.image-net.org/challenges/LSVRC/2012 .

[2] Project webpage, 2013. http://users.cecs.anu.edu.au/~xlx/proj/tagnet .

[3] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. CHI '07, pages 971–980, 2007.

[4] K. Atkinson. Official 12Dicts package. http://wordlist.sourceforge.net , retrieved March 2012.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.

[6] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. WSDM '11, pages 635–644, 2011.

[7] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[8] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569. IEEE, 2012.

[9] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? CIKM '08, pages 193–202, 2008.

[10] N. Chen, Q.-Y. Zhou, and V. Prasanna. Understanding web images by object relation network. WWW '12, pages 291–300, 2012.

[11] M. Choi, A. Torralba, and A. Willsky. A tree-based context model for object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(2), 2012.

[12] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *ACM CIVR*, 2009.

[13] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255. Ieee, 2009.

[14] J. Deng, S. Satheesh, A. Berg, and L. Fei-Fei. Fast and balanced: Efficient label tree learning for large scale object recognition. In *NIPS*, 2011.

[15] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *CVPR*, pages 1777–1784. IEEE, 2011.

[16] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, I. Hal Daumé, and A. Berg. Detecting visual text. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2012.

[17] C. Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243, 2010.

[18] G. H. Golub and C. D. Meyer, Jr. Using the qr factorization and group inversion to compute, differentiate ,and estimate the sensitivity of stationary probabilities for markov chains. *SIAM J. Algebraic Discrete Methods*, 7(2):273–281, Apr. 1986.

[19] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *ACM Trans. Knowl. Discov. Data*, 5(4):21:1–21:37, Feb. 2012.

[20] T. L. Griffiths, M. Steyvers, and A. Firl. Google and the mind predicting fluency with pagerank. *Psychological Science*, 18(12):1069–1076, 2007.

[21] C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, pages 27–29, 2007.

[22] J.-W. Jeong, X.-J. Wang, and D.-H. Lee. Towards measuring the visualness of a concept. CIKM '12, pages 2415–2418, 2012.

[23] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM Multimedia*, pages 631–640, 2007.

[24] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[25] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[26] A. Langville and C. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2009.

[27] L.-j. Li, H. Su, E. P. Xing, and L. Fei-fei. Object Bank : A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In *NIPS*, pages 1–9, 2010.

[28] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Trans. Multi.*, 11(7):1310–1322, Nov. 2009.

[29] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. WWW '09, pages 351–360, 2009.

[30] T. Malisiewicz and A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009.

[31] T. Mitchell, S. Shinkareva, A. Carlson, K. Chang, V. Malave, R. Mason, and M. Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

[32] J. Morales and J. Nocedal. Remark on "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". *ACM Transactions on Mathematical Software (TOMS)*, 38(1):7, 2011.

[33] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 13(3):86–91, 2006.

[34] S. Overell, B. Sigurbjörnsson, and R. van Zwol. Classifying tags using open content resources. WSDM, 2009.

[35] K. B. Petersen and M. S. Pedersen. The matrix cookbook, oct 2008. Version 20081110.

[36] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative multi-label video annotation. In *ACM Multimedia*, pages 17–26, 2007.

[37] G.-J. Qi, C. Aggarwal, and T. Huang. Towards semantic knowledge propagation from text corpus to web images. WWW '11, pages 297–306, 2011.

[38] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1575–1589, 2007.

[39] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. SIGIR, pages 103–110. ACM, 2007.

[40] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. WWW'08, pages 327–336. ACM, 2008.

[41] D. Stern, R. Herbrich, and T. Graepel. Matchbox: large scale online bayesian recommendations. In *Proc. World wide web*, pages 111–120. ACM, 2009.

[42] F. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *World Wide Web*, pages 697–706. ACM, 2007.

[43] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM Multimedia*, pages 223–232, 2009.

[44] H. Tong, C. Faloutsos, and J. Pan. Fast random walk with restart and its applications. In *ICDM*. IEEE, 2006.

[45] K. Van De Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Trans.*, 32(9):1582–1596, 2010.

[46] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Visual tag dictionary: interpreting tags with visual words. In *Workshop on Web-scale multimedia corpus*, 2009.

[47] X.-J. Wang, Z. Xu, L. Zhang, C. Liu, and Y. Rui. Towards indexing representative images on the web. In *ACM Multimedia*, pages 1229–1238, 2012.

[48] K. Q. Weinberger, M. Slaney, and R. Van Zwol. Resolving tag ambiguity. In *ACM Multimedia*, pages 111–120, 2008.

[49] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proc. IJCAI*, pages 2764–2770. AAAI Press, 2011.

[50] K. Yanai and K. Barnard. Image region entropy: a measure of "visualness" of web images associated with one concept. In *Proc. ACM Multimedia*, pages 419–422, 2005.