Learning from the Uncertain

Leveraging Social Communities to generate reliable Training Data for Visual Concept Detection Tasks

Christian Hentschel
Hasso Plattner Institute for Software Systems Engineering
Potsdam, Germany
christian.hentschel@hpi.de

Harald Sack
Hasso Plattner Institute for Software Systems Engineering
Potsdam, Germany
harald.sack@hpi.de

ABSTRACT

Recent advances for visual concept detection based on deep convolutional neural networks have only been successful because of the availability of huge training datasets provided by benchmarking initiatives such as ImageNet. Assembly of reliably annotated training data still is a largely manual effort and can only be approached efficiently as crowdworking tasks. On the other hand, user generated photos and annotations are available at almost no costs in social photo communities such as Flickr. Leveraging the information available in these communities may help to extend existing datasets as well as to create new ones for completely different classification scenarios. However, user generated annotations of photos are known to be incomplete, subjective and do not necessarily relate to the depicted content. In this paper, we therefore present an approach to reliably identify photos relevant for a given visual concept category. We have downloaded additional metadata for 1 million Flickr images and have trained a language model based on user generated annotations. Relevance estimation is based on accordance of an image's annotation data with our language model and on subsequent visual re-ranking. Experimental results demonstrate the potential of the proposed method – comparison with a baseline approach based on single tag matching shows significant improvements.

CCS Concepts

•**Information systems** → **Image search;** *Data mining;*

Keywords

social image retrieval, relevance estimation, language model, visual re-ranking

i-KNOW '15, October 21 - 23, 2015, Graz, Austria

ISBN 978-1-4503-3721-2/15/10...$15.00

DOI: http://dx.doi.org/10.1145/2809563.2809587

1. INTRODUCTION

Visual concept detection refers to the ability of learning visual categories in order to automatically identify new, unseen instances of these categories. Typically, this task is approached as a supervised machine learning task: by using a reliably annotated dataset of example images separated into the categories the system should be able to recognize, a machine learns distinguishing features of the individual categories. Recently, approaches based on deep convolutional neural networks (CNN) have significantly improved over previous methods in terms of achieved classification precision. In fact, these systems have even reported to achieve better-than-human recognition rates in certain benchmarks [2]. Their success must firstly be attributed to the availability of massively parallel hardware that supports training very large and very deep neural networks in a reasonable amount of time. The larger and deeper a network becomes, the more parameters need to be optimized which makes training prone to overfitting when only a limited amount of labeled training data is available. Hence, a second and equally important factor for the success of deep neural networks was the availability of vast amounts of training data. Benchmarking initiatives such as the ImageNet[1] – a database of currently more than 14 million images categorized into more then 21,000 classes in a huge crowd sourcing effort – are usually applied as data sources to train CNNs. The declared goal of the ImageNet initiative is to support 40,000 categories each covered by 10,000 individual images each of which evaluated as relevant by the majority of 10 human annotators [19]. Human labeling speed is estimated by 2 images per second and thus the overall human annotation time required is calculated as more than 63 years. Clearly, an effort as such will be difficult if not impossible to repeat but even the extension of existing categories or adding new ones represents a major challenge if one cannot rely on an armada of crowd workers.

On the other hand the World Wide Web provides huge data sources of annotated visual content almost for free. Photo sharing platforms such as Flickrhost billions of user-generated images[2]. The community aspect has motivated millions of users to manually annotate their images with descriptive metadata such as image titles, tags and descriptions in order to increase the visibility of their photos or

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

[2]Flickr reports to host more than 6 billion photos: http://blog.flickr.net/en/2011/08/04/6000000000/

share their content with other users. Being able to exploit these information as training data not only would enable enlargement of existing datasets and categories by additional images. Considering the potentially unlimited vocabulary represented in user annotations a huge variety of additional categories could instantly be made available at almost no costs (even across a multitude of different languages).

However, a major drawback of these user annotations arising from the uncontrolled environment in which they are generated is that they need to be considered incomplete, highly subjective and not necessarily related to the visual content of the respective photo ([4, 12, 14]). This contrasts sharply with highly reliable image annotations required for learning visual concept classifiers and aimed at by initiatives such as ImageNet. Incompleteness – meaning that not all photos which depict a specific visual concept are actually annotated with a textual label identifying that concept – is usually of minor importance due to the potentially unlimited amount of annotated data available in photo communities. Subjective annotations and annotations with missing relevance to the depicted content, however, pose a major challenge when trying to retrieve images suitable to train a specific visual concept classifier.

In this paper, we present an approach to automatically select photos – which are considered relevant to a given visual concept label by a majority of users – from a large collection of publicly available Flickr images. Our key assumption is that the majority of all users in a community shares a common (and thereby objective) interpretation of a visual concept. This is reflected by the fact that users in general employ a highly similar vocabulary to describe a visual concept whereas overly subjective annotations will exhibit a comparatively low inter-annotator agreement.

In order to verify our assumptions, we have downloaded authoritative and social metadata for almost 1 million Flickr images. The metadata corpus will be released upon publication of this paper. We have trained a language model on the entire corpus which allows us to extract contextually similar terms for a given query term, e.g. a visual concept label. Using these terms, we extract candidate images for a specific visual concept based on their textual similarity with our language model. In order to further increase relevance we apply a visual re-ranking of the top ranked candidate images using deep feature representations. Evaluation results, reported Average Precision scores, prove the validity of our approach.

This paper is structured as follows: In Section 2 we briefly review the related work. The learned language model as well as our approach for visual re-ranking is presented Section 3 where we also introduce the dataset that we have assembled as well as our experimental setup. We present and discuss evaluation results in Section 4. Finally, Section 5 concludes the paper and gives a brief outlook to future improvements.

2. RELATED WORK

The availability of large photo communities or web-based image collections as well as user generated annotations has triggered many researchers to exploit these information in completely different scenarios. Next to the obvious research goal to increase relevance for search and retrieval within these collections research scenarios include retrieval of photos related to a tourist attraction [17], landmark recognition [9] as well as automatic image labeling [21], which is also our declared goal. All approaches denote the comparatively large amount of noise within community based annotations and propose different strategies to increase the relevance of the retrieved results. Typically, these approaches rely on the textual annotations, compute visual similarities between images or employ a combination of both strategies.

Text-based retrieval typically tries to increase relevance by extending the initial query. In [11] a dictionary lookup reveals additional tags which are expected to be suited for describing visual content. Later, a neighbor voting algorithm is applied to the identified tags. However, a dictionary usually does not describe visual content in a similar way it is described by users of a social community. Related tags indicated by resources such as external knowledge bases like DBpedia or taxonomies as e.g. WordNet usually do not provide enough relevant information for user annotation relationship modeling. In [5] we used DBpedia to model inter-tag relationships and found out that relevant information is usually not available as direct link pattern. Here, we therefore use a language model trained on the corpus of available user annotations itself.

Approaches based on visual similarity estimation assume that all images relevant to a specific visual concept tend to have similar visual patterns such as color and texture. Typically, density estimation methods are used to find the dominant visual pattern in the form of clustering [16, 6]. The typically large feature dimensionality as well as insufficient number of samples usually, however, renders density estimation inaccurate and makes similarity computation expensive [12]. Our approach uses feature representations based on deep convolutional neural networks that are known to generate compact visual features. Furthermore, to decrease computational complexity, we compute visual similarities only for a subset of candidate images returned by prior textual ranking. In [20] the authors exploit visual consistency to quantify the representativeness of Flickr images with respect to a given tag. Relevance in our approach is estimated by a language model directly trained from the textual annotations of millions of users and thereby considered to be objective. Visual consistency is further increased by re-ranking candidate images based on visual features.

In order to jointly exploit visual and semantic consistency, in [13] tag relevance estimation is approached as a semi-supervised multi-label learning problem. The authors in [24] attempt to solve the same problem by decomposing an image tag co-occurrence matrix and Yang et al. [22] present a framework, which simultaneously refines the noisy tags and learns image classifiers. In [12] the authors consider the task of estimating the relevance of a Flickr tag for the depicted visual concept. Based on visual similarity of the k nearest neighbors of an image, co-occurring tags are ranked higher. While this helps to boost relevant tags within the respective neighborhood, it does not help to solve the problem of selecting images as training examples. Contextual information is ignored and thus different meanings of the same tag cannot be distinguished. Our approach uses contextual information in order to select images matching the most likely meaning of an annotation within a photo community.

3. RELEVANT IMAGE RETRIEVAL

In this sections we present our approach to retrieve photos relevant for a given visual concept from a large collection of 1 million Flickr images. We start by presenting the collection

that we have used and extended throughout our experiments as well as some statistics on the downloaded metadata. Second, we detail our approach for text-based image retrieval using a language model trained on the available metadata. We further describe our visual re-ranking step achieved by computing image similarity using deep visual feature representations. Finally, we present our experimental setup in order to evaluate the relevance of the retrieved photos with regard to potential use as traning data.

3.1 Dataset

The Flickr platform provides a public API[3] to query their database and has already been used by many research activities in the past as it has become relatively easy to access a huge amount of photos and metadata. Various official datasets and benchmarking initiatives in content based image classification and image retrieval have made use of images downloaded from the Flickr community. As an example, the MIRFLICKR-1M collection was published in [7] and consists of 1 million images crawled from Flickr. The selection of images has been made based on the Flickr *interestingness* score – a measure that aggregates factors such as clickthrough rate, user comments as well as users selecting an image as favorite. The data has been made available under a Creative Commons Attribution license (meaning that each image can be used as long as the photographer is credited for the original creation) and includes EXIF metadata as well as raw user tag data. Although the authors attribute the value of additional descriptive and social metadata such as title and description as well as information about the owner's social network, the dataset unfortunately does not contain these data. Fortunately, the provided license information for each individual image permits to download these information based on the unique photo id using the Flickr API.

Being a social community, the information stored by Flickr are not static i.e. photos as well as accompanying metadata undergo changes throughout time. Users may delete existing photos and add, remove or change metadata, aggregate photos in public groups and private sets or comment on the photos of other users. The data downloaded, hence, should be considered as a snapshot taken at a specific time and very likely to be different at any time in the future. Consequently, since 5 years elapsed between the initial assembly of the MIRFLICKR-1M collection and our most recent extension, some of the photos have been removed by their owners meanwhile and no metadata can be retrieved. For the majority of 90.2% of the entire collection (902,672 photos), however, we were able to assemble additional information. Table 1 gives a detailed overview of the amount of available metadata.

In addition to the metadata content itself we have also downloaded the unique id of the responsible user (available for comments, tags, notes, and photo ownership) that will make it more easy to identify connections between photos within the social network. We make all metadata publicly available as individual JSON files[4] to be used in future research projects.

Most of the downloaded metadata is stored in form of unstructured free text (title, description, user comments) or single word labels (tags). In [4] the authors analyze different usage pattern of collaborative tagging systems which can be extended to annotations in general and likewise hold for photo communities such as Flickr. The overwhelming majority of user annotations is reported to be used to identify the topic or content of the annotated data – an important prerequisite when aiming to match described and depicted content in photos. In [5] we have analyzed Flickr photos and user generated tags for relatedness. We've found out that annotations next to identifying the depicted content may also be used with an organizational or viewpoint defining purpose. Thus, even when an annotation explicitly mentions a visual concept this does not necessarily mean it is actually depicted. An algorithm that selects photos for usage as training data based on textual annotations should therefore be able to identify these photos as not being relevant for the respective visual concept. In this paper we define a photo being relevant for a given visual concept if it depicts a clearly-visible version of the scene or object without any major occlusion.

Table 1: Amount of available metadata for photos of the MIRFLICKR-1M collection

Total no. of photos	1,000,000
w/ title	864,081
w/ description	607,663
w/ tags	858,918
w/ EXIF data	688,294
w/ geo information	282,091
w/ album allocation	760,702
w/ user comments	851,174
w/ notes	102,252
w/ group allocation	740,263

3.2 A Community Language Model

For the aforementioned reasons, selecting photos solely based on the usage of the visual concept term within the annotations will likely fail to provide relevant photos. However, when considering the annotation context we assume that the majority of all users will use a similar vocabulary to describe the content. As an example, a photo depicting a sunset in many cases will also contain annotations such as "sea", "ocean", "clouds". When extending the query for "sunset" by these additional terms the retrieved photos will tend to exhibit higher relevance to the initial concept. Yet, manual creation of an extended query vocabulary per visual concept is prone to errors, subjective and most likely does not capture every relevant term.

In [5] we therefore used tag disambiguation and link analysis to automatically extract related tags from a knowledge base (DBpedia). Based on the number of tags of an image that have a direct link to the visual concept within the knowledge base we have estimated how strongly a tag set of a given image is related semantically to the respective concept and thus, how much related the image is. However, we have found out that frequently co-occurring tags such as "bridge" and "river" do not exhibit direct links in the knowledge base and are thus much harder to identify.

[3]The Flickr API: https://www.flickr.com/services/api/

[4]The MIRFLICKR-1M s16a extension: http://s16a.org/mirflickr

In this paper, we therefore decided to learn annotation relationships based on contextual similarity immediately from the metadata corpus itself. This not only reduces a potentially error-prone manual query extension but also extracts additional terms based on the community users' applied vocabulary.

The authors in [15] present a neural network based approach to learn vector representations of single words, accordingly named *word2vec*. Training is performed in a completely unsupervised fashion – given a sufficiently large corpus, such as textual image annotations. A trained word2vec model allows to make highly accurate predictions about a word's meaning based on past contextual appearances. The authors suggest two architectures to learn the underlying word representation: Continuous Bag-of-Words (CBOW) and continuous skip-gram. Both architectures define a way of how to create labels for training word representations in an unsupervised scenario. While CBOW predicts a word given its surrounding words (or context), skip-gram predicts the context given a specific word. In both cases, the window size parameter defines the size of the respective context. According to the authors, training of the skip-gram model is slightly slower but the architecture is better suited to represent infrequent words. The output of a word2vec model is a vocabulary of all learned words and their respective vector representations. These can be used to compute the cosine similarity of words. For a more detailed description and comparison of both algorithms we refer the reader to [15].

For our experiments, we have used the skip-gram implementation as provided in the gensim python package [18]. By training a word2vec model on the textual user annotations we enable extraction of similar terms given a visual concept label according to the language used by Flickr users. As an example, we have extracted the 10 most similar terms for the concept 'sunset':

sunset	dusk, sundown, sun, twilight, sunrise, cloud, silhouette, settingsun, nightfall, sky

Apparently, our initial assumption of the model being able to extract community specific related terms holds: while terms such as 'sun', 'sunrise', 'cloud' and 'sky' could have also been manually selected as plausible query extension, the artificial term 'settingsun' can be only learned from the data itself.

3.3 Visual Re-ranking

As discussed in Sect. 1 we aim at increasing relevance by re-ranking candidate images based on their visual similarity. It has been shown that features extracted from the activation of a deep convolutional neural network which has been trained to separate individual visual concept categories on a large dataset can be reused and adapted to novel classification tasks [3, 23]. It has been further shown that these novel tasks may differ from the original training scenario and that deep feature encodings significantly outperform any previously presented "shallow" encodings (e.g. Bag-of-Visual-Words, Fisher encoding etc., for a comparison of shallow and deep encodings in generic image classification tasks, see [1]). In [23] the authors explored how discriminative the features in each layer of a CNN model trained on one dataset are for classifying a different dataset. This is done by forward propagating test images through a varying number of layers of the trained model and training a linear SVM classifier. The results showed that with increasing number of layers, the classification results also improved supporting the notion that the first layers in a neural network learn "low-level" features, whereas the latter layers learn semantic or "high-level" features.

In order to obtain compact visual feature representations we make use of these findings by taking a deep convolutional neural network pre-trained on the ILSVRC-2012 dataset[5]. The model is provided as part of Caffe CNN implementation [8] and extends from the successful architecture presented in [10]. Specifically, compared to the original architecture the authors use slightly different data augmentation techniques and switch the order of the pooling and normalization layer achieving a slightly reduced classification time. For a detailed discussion of the CNN architecture and training protocol we refer the reader to [10].

In our experiments, we have used the vector of activities of the penultimate, fully-connected (seventh) layer (fc7) as feature descriptors, obtaining a 4,096 dimensional descriptor vector per image. We extracted the fc7-features for all images in the MIRFLICKR-1M collection. Using an NVIDIA Tesla K20 GPU, feature extraction took about 3 hours. We publish the extracted features as an additional extension of the collection[4].

Now we can determine the similarity of two candidate images i_1, i_2 by computing the cosine similarity of their respective layer-7 activity representations $fc7$:

$$k(i_1, i_2) = \frac{fc7(i_1)fc7(i_2)^T}{\|fc7(i_1)\|\|fc7(i_2)\|}. \quad (1)$$

3.4 Experimental Setup

We test our approach on 10 selected visual concept categories. These categories comprise 8 object-level concepts ('airplane', 'bicycle', 'boat', 'bridge', 'car', 'dog', 'flower', and 'tiger') and 2 scene-level concepts ('beach' and 'mountain') and follow the categories chosen by the authors in [12].

At this stage we train the language model using tag-based annotations only. While we plan to extend to other annotations, tags provide the immediate advantage of being single word terms whereas titles, descriptions and user comments for example may appear as HTML encoded full-text strings that makes parsing and tokenization more error-prone.

We preprocess all tags by running lemmatization and stopword removal. Currently, we focus on English language only meaning that any other language is ignored for preprocessing. Analysis of the downloaded metadata corpus shows that users add an average of $|tags| = 12$ unique tags per photo ($\mu = 12.429, \sigma = 10.235$). We therefore set the window size for training our word2vec model to $w = |tags|/2 = 6$ words and ignore all words with a total frequency of $f < 5$. We train 300-dimensional feature vectors on the tag data and compute the $k = 20$ most similar terms for each visual concept label. Table 2 shows the concept labels and the (top-10) most similar terms according to our model.

The advantages of our model become easily visible: not only are relevant synonyms extracted (e.g. *airplane*: [air-

[5]ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012): http://www.image-net.org/challenges/LSVRC/2012/

craft, aeroplane, plane]) but also terms obviously *related* to the concept are selected (e.g. *beach*: [sand, ocean, shore]). Furthermore, frequent instances (e.g. *flower*: [dahlia, spiderwort, hyacinthaceae], *car*: [ford]) are extracted as well as translations of the original concept into different languages (e.g. *dog*: [chien]). Similarly, the model captures sub- and superclass relationships (e.g. *boat*: [fisherboat, sailboat, sailingships] and *tiger*: [flickrbigcats]) automatically from the dataset without having to extract them from an external knowledge base.

Based on these 20 most similar terms, we construct an extended query including the visual concept label. For each concept we then select those photos from the collection that best match the extended query assuming that images ranked higher are more likely to be relevant candidate images. We therefore rank images based on the number of query terms found in the respective tagset. Table 3 shows the top ranked images including the corresponding list of user generated tags. Tags that match our extended vocabulary are typed in boldface.

Visual re-ranking is applied to further increase the relevance of the top ranked images. We assume that the highest ranked image based on our extended query exhibits a high relevance for the visual concept and therefore re-rank the remaining images based on visual similarity to the top candidate. We compute the cosine similarity on the extracted deep feature representations as presented in Sect. 3.3.

4. RESULTS

This section presents the results obtained for each step in the retrieval process. We compare our approach to a simple baseline algorithm (thus referred to as *baseline* hereafter), which selects photos based on whether or not the tagset contains the visual concept label (i.e. without any query extension). The number of candidate images based on this simple approach is considerably large. In order to evaluate the accuracy of the baseline method, we randomly sample $n = 200$ images for each visual concept category.

In order to evaluate a potential gain in accuracy by the individual steps, we have separately evaluated retrieval results based on the learned language model as well as based on additional visual re-ranking. Since both approaches generate ranked result set, we take the top $n = 200$ ranked candidates for evaluation. We manually assess the relevance of candidates for all three approaches following our definition in Section 3.1: A photo is considered relevant if it clearly depicts the scene or object without any major occlusion. Evaluation results are reported as average precision scores corresponding to the area under the precision-recall-curve. Given a ranked list L with length n, average precision is defined as:

$$AP = \frac{1}{R} \sum_{k=1}^{n} \frac{R_k}{k} * rel(k) \quad (2)$$

where R is the number of relevant photos in L, R_k the number of relevant images among the top k ranked instances, $rel(k) = 1$ is an indicator function equaling 1 if the photo at rank k is relevant, 0 otherwise.

The results reported in Table 4 show the superiority of the proposed method. In general, the approach based on visual re-ranking of the results obtained from the trained language model outperforms the baseline approach as well as ranking based on textual features only. There are two major exceptions: While the results obtained for the category "airplane" based on the language model clearly outperform the baseline

Table 2: Visual concept labels and most similar terms according to skip-gram community language model

concept	similar terms
airplane	aircraft (0.90), aviation (0.88), aeroplane (0.86), plane (0.85), jet (0.85), airliner (0.84), jetliner (0.82), cockpit (0.81), regionaljet (0.78), planespotting (0.77)
beach	sand (0.78), ocean (0.70), shore (0.70), surf (0.69), wave (0.68), sea (0.67), zwemmen (0.64), kontikiinn (0.62), lowtide (0.62), capehenlopenstatepark (0.61)
bicycle	bike (0.88), cycle (0.88), cycling (0.85), citycycling (0.77), cyclist (0.77), bikelanes (0.77), bikelane (0.76), ridealong (0.76), citycycle (0.76), biking (0.76)
boat	sailing (0.79), ship (0.79), sail (0.77), moored (0.74), dock (0.74), yacht (0.73), fishingboats (0.73), sailboat (0.73), sailingship (0.72), port (0.72)
bridge	suspensionbridge (0.62), river (0.61), suspension (0.56), footbridge (0.56), swingbridge (0.53), building (0.52), brigde (0.52), riverhumber (0.51), barge (0.51), reka (0.51)
car	automobile (0.79), auto (0.76), sportscar (0.76), convertible (0.74), coupe (0.74), luxurycar (0.73), 6car (0.72), sedan (0.72), customcar (0.71), ford (0.71)
dog	puppy (0.89), canine (0.81), mutt (0.78), k9 (0.77), terrier (0.77), chien (0.76), interestingdogsposes (0.76), retriever (0.75), doggy (0.75), pup (0.74)
flower	bloom (0.74), daisy (0.72), flora (0.71), dahlia (0.70), spiderwort (0.69), hyacinthaceae (0.69), columbine (0.68), petal (0.68), flowercloseup (0.68), coneflower (0.67)
mountain	peak (0.73), hiking (0.69), mountainrange (0.68), snowcapped (0.68), valley (0.67), glacier (0.66), mountaineering (0.65), alpine (0.65), treck (0.64), gipfel (0.63)
tiger	flickrbigcats (0.61), pantheratigris (0.59), amurtiger (0.57), siberiantiger (0.56), cub (0.56), whitetiger (0.55), tigercub (0.55), sumatrantiger (0.55), bengaltiger (0.55), eagle (0.54)

Table 3: Top-ranked candidate images according to skip-gram language model. User tags that match our extended vocabulary are typed in boldface.

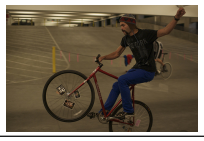
concept	user tags	top ranked image
airplane	passenger, vliegtuig, aéroport, jetplane, traveller, **airliner**, lesavions, economysection, traveler, vacation, fuselage, motor, **jetliner**, transportation, **airplane**, legroom, transport, sky, passengerjet, **jet**, travel, flying, **avion**, airport, inflight, rudder, flugzeug, tail, bin, ptvs, aerial, cabin, passengerplane, aircraftpicture, schipholairport, flap, nederland, avião, amsterdam, landinggear, **cockpit**, luggagebins, **aircraft**, **plane**, **aviation**, seat, economyclass, airship, aisle, netherlands, nosegear, holland, luggage, aircraftcabin, engine, aeroport, ptv, **aeroplane**, aeroplano, wing, paysbas	
beach	**sea**, blackandwhite, ca, reflected, bright, seascape, seashore, coastal, monochrome, pacificocean, touristdestination, placeofinterest, usa, torreypines, travel, **shore**, black, weather, white, january08, **beach**, **surf**, evening, colorless, grayscale, sunshine, sunny, stonematerial, **coast**, reflect, bw, california, **coastline**, **wave**, tourism, glow, water, touristattraction, reflection, sandiego, light, **seaside**, **ocean**, **sand**, reflecting, sunset, rock, beachculture, blackwhite	
bicycle	**cyclist**, 2wheelsgood, whatswheeliegood, **cycling**, cog, bicyclist, bicycling, fixiewhippingood, **bike**, **fixie**, **bicycle**, **fixed**, **biking**, **fixedgear**	
boat	vell, marinaportvell, engaged, boardwalk, street, vatalonia, **port**, **boat**, **sailboat**, portvell, engagement, **sail**, espana, band, marina, pier, spain, engage, reflection, **harbor**, barcelona, **mast**, tonemapped, 400d	
bridge	beautiful, **bridge**, lavender, high, city, wire, luz, sky, greenville, **suspension**, white, greenvillesc, park, dark, southcarolina, fall, cityscape, line, upstate, light, tree, **suspensionbridge**, night, sc, **river**	
car	harney, **ford**, illinois, myoldpostcards, leeannharney, il, owner, **coupe**, chromeengine, backend, taillight, **automobile**, route66, custom, tail, motorvehicle, vintagecar, fomoco, international, fin, collectiblecar, 2012, custombuilt, september21232012, **auto**, dougthompson, motherroadfestival, 1950, 9212312, **convertible**, fordmotorcompany, **classiccar**, 2door, worldcars, **car**, ghostflames, vonliski, rearend, sidepipes, antiquecar, springfield, frankharney, deluxe, carsonconvertibletop, **oldcar**	
dog	k10d, cute, pentax, cane, **labrador**, golden, sweet, canon, **canine**, retreiver, portrait, funny, eye, winter, pet, lake, perro, animal, **puppy**, happy, brown, play, abigfave, labradorretriever, lab, bw, sigma1020mm, **pup**, chala, **retriever**, **dog**, nose, fun, **chien**	
flower	stamen, **plant**, **flower**, tamron90mmlens, garden, 360pxflash, blossom, petal, 000afflash, pentaglottissempervirens, sonya100, 4, **bloom**, dioptrelenses, kenkouniplus25extensiontube	
mountain	**mountain**, september, canada, provincial, chilliwack, mountjudgehoway, snow, pacificranges, staveriver, coastmountains, granite, bc, park, judgehoway, judge, howay, **alpine**, double, mount, thejudge, **summit**, britishcolumbia, fraservalley, **peak**	
tiger	tigris, zurich, panthera, felid, **flickrbigcats**, impressedbeauty, portrait, close, face, openmouth, **siberiantiger**, feline, young, zoo, **bigcat**, standing, head, big, kitty, **tiger**, wild, closeup, stripe, schweiz, **coto**, striped, **wildcat**, nikon, zürich, **amurtiger**, goldstaraward, tigre, cat, d300, **pantheratigris**, switzerland	

Table 4: Comparison of proposed approaches for relevant image retrieval. Skip-gram is our approach based on the proposed community language model only. Skip-gram+vr denotes results obtained after additional visual re-ranking. Reported scores are average precision. Best results are marked in boldface.

concept	baseline	skip-gram	skip-gram+vr
airplane	0.457	**0.797**	0.237
beach	0.512	0.615	**0.812**
bicycle	0.476	0.850	**0.993**
boat	0.549	0.712	**0.936**
bridge	0.450	**0.611**	0.513
car	**0.621**	0.597	0.162
dog	0.758	0.885	**0.953**
flower	0.828	0.941	**0.980**
mountain	0.619	0.863	**0.977**
tiger	0.551	0.803	**0.959**
Mean AP	0.582	**0.765**	0.752

approach, we see a significant drop in the reported average precision when applying visual re-ranking. This is likewise true for "car" where the baseline approach even outperforms the textual model by 2%. Considering the top ranked image used as seed image for visual re-ranking for both classes (see Table 3) we see that the image ranked highest according to our language model for the category "airplane" actually depicts an airport (although the number of found vocabulary tags indicate a high relevance for the "airplane" category). Visual re-ranking is therefore based on an airport image and fails to capture essential features of airplanes. Similarly, the highest ranked image for the category "car" actually depicts the rear light of an old car. Both misclassifications heavily decrease the achieved AP score and thus also affect the mean average precision score which is therefore slightly worse for the combination of language model and visual re-ranking. To avoid this in future, we plan to include more than just the top ranked photo for computation of visual similarities. An option that we consider is to train a single-class classifier based on the top-n highest ranked candidates according to our language model.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an approach to retrieve relevant images from a large corpus of Flickr images. We have started by extending the MIRFLICKR-1M collection by additional user generated annotations data which we make publicly available. Using the tagset of all images we have trained a word2vec based community language model. Our method starts by retrieving the top-n candidate images according to the language model and we further refine results based on computation of deep visual feature representations. Reported evaluation results prove the superiority of our approach over a baseline method that retrieves images based on exact tag matching.

The work presented here is only a first step towards exploitation of community photo data for visual concept classification. Currently, image relevance estimation for a given concept is based on manual assessment of a single user. An image is considered relevant if it clearly depicts the scene or object without any major occlusion. As a matter of fact, this implies a strong bias towards the respective evaluator. In future, we will therefore consider relevance estimations of different evaluators and use inter-annotator agreement in order to obtain more objective assessments.

As discussed we aim to include further annotation data such as title, description and Flickr group information into our language model. Second, we aim to optimize parameters such as the number k most similar terms used to extend our initial query. Furthermore, we will train a classifier that considers the top-n candidate images to improve visual re-ranking. Finally, we will test the retrieved results in classification scenarios, i.e. we will evaluate the performance achieved by visual concept classifiers when trained on photos returned using our methods.

6. REFERENCES

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In M. Valstar, A. French, and T. Pridmore, editors, *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[2] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32:333–338, 2012.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *International Conference on Machine Learning*, pages 647–655, 2014.

[4] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems, 2006.

[5] C. Hentschel, H. Sack, and N. Steinmetz. Cross-Dataset Learning of Visual Concepts. In A. Nürnberger, S. Stober, B. Larsen, and M. Detyniecki, editors, *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, volume 8382, pages 87–101. Springer International Publishing, 2013.

[6] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 35–44, New York, NY, USA, 2006. ACM.

[7] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection. In *Proceedings of the international conference on Multimedia information retrieval - MIR '10*, page 527, New York, New York, USA, 2010. ACM Press.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the ACM International Conference on Multimedia - MM '14*, pages 675–678, 2014.

[9] 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 *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA '07, pages 631–640, New York, NY, USA, 2007. ACM.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[11] S. Lee, W. De Neve, and Y. M. Ro. Image tag refinement along the 'what' dimension using tag categorization and neighbor voting. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 48–53, July 2010.

[12] X. Li, C. G. M. Snoek, and M. Worring. Learning Tag Relevance by Neighbor Voting for Social Image Retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 180–187, New York, NY, USA, 2008. ACM.

[13] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 491–500, New York, NY, USA, 2010. ACM.

[14] Matusiak and K. K. Towards user-centered indexing in digital image collections, 2006.

[15] T. Mikolov, G. Corrado, K. Chen, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, 2013.

[16] G. Park, Y. Baek, and H.-K. Lee. Majority based ranking approach in web image retrieval. In E. Bakker, M. Lew, T. Huang, N. Sebe, and X. Zhou, editors, *Image and Video Retrieval*, volume 2728 of *Lecture Notes in Computer Science*, pages 111–120. Springer Berlin Heidelberg, 2003.

[17] A. Popescu and G. Grefenstette. Deducing trip related information from flickr. In *Proceedings of the 18th international conference on World wide web*, pages 1183–1184. ACM, 2009.

[18] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

[20] A. Sun and S. S. Bhowmick. Quantifying tag representativeness of visual content of social images. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 471–480, 2010.

[21] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1919–1932, Nov. 2008.

[22] Y. Yang, Y. Gao, H. Zhang, J. Shao, and T.-S. Chua. Image tagging with social assistance. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 81:81–81:88, New York, NY, USA, 2014. ACM.

[23] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014, 13th European Conference*, volume 8689, pages 818–833. Springer International Publishing, 2014.

[24] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 461–470, New York, NY, USA, 2010. ACM.