Delving Deep into Personal Photo and Video Search

Lu Jiang[1]*, Yannis Kalantidis[2], Liangliang Cao[2], Sachin Farfade[2], Jiliang Tang[3], Alexander G. Hauptmann[1]

[1] Carnegie Mellon University, [2] Yahoo Research, [3]Michigan State University

{lujiang, alex}@cs.cmu.edu, liangliang.cao@gmail.com, ykalant@image.ntua.gr, fsachin@yahoo-inc.com, tangjili@msu.edu

ABSTRACT

The ubiquity of mobile devices and cloud services has led to an unprecedented growth of online personal photo and video collections. Due to the scarcity of personal media search log data, research to date has mainly focused on searching images and videos on the web. However, in order to manage the exploding amount of personal photos and videos, we raise a fundamental question: what are the differences and similarities when users search their own photos versus the photos on the web? To the best of our knowledge, this paper is the first to study personal media search using large-scale real-world search logs. We analyze different types of search sessions mined from Flickr search logs and discover a number of interesting characteristics of personal media search in terms of information needs and click behaviors. The insightful observations will not only be instrumental in guiding future personal media search methods, but also benefit related tasks such as personal photo browsing and recommendation. Our findings suggest there is a significant gap between personal queries and automatically detected concepts, which is responsible for the low accuracy of many personal media search queries. To bridge the gap, we propose the deep query understanding model to learn a mapping from the personal queries to the concepts in the clicked photos. Experimental results verify the efficacy of the proposed method in improving personal media search, where the proposed method consistently outperforms baseline methods.

CCS Concepts

•**Information systems** → **Image search;** *Multimedia and multimodal retrieval;* •**Computing methodologies** → **Neural networks;**

Keywords

Content-based Image Search; Search Log Analysis; Deep Learning; Recurrent Neural Networks

*This work was done when the first author was on an internship at Yahoo! Research.

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

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

© 2017 ACM. ISBN 978-1-4503-4675-7/17/02…$15.00

DOI: http://dx.doi.org/10.1145/3018661.3018736

1. INTRODUCTION

Personal photo and video data are being accumulated at an unprecedented speed. For example, 14 petabyte of personal photos and videos were uploaded to Google Photo[1] by 200 million users in 2015 [13], while tremendous amount of personal photos and videos are also being uploaded to Flickr[2] every day. With personal media[3] now ubiquitous online and given the added sentimental value such data carry, searching them efficiently and effectively is becoming increasingly important as user's memories accumulate over time. Personal media, stored in mobile devices or online cloud, are becoming mirrors to one's everyday life and past memories, capturing significant or cherishable moments, *e.g.* , wedding ceremonies or birthday parties.

With the speed that media gets created everyday, manually annotating personal photo archives is practically infeasible. This is a situation comparable to the days in the late 1990s, when people usually got lost in the rising sea of web pages, now they are overwhelmed by the vast amounts of personal media data but lack tools to find desired information. In the absence of textual metadata, searching can only be achieved via other automatically generated metadata (*e.g.* timestamps and location provided by the recording device). In recent years, deep learning based image classification [30] has been employed for media auto-tagging; photos can automatically be annotated with generic concepts, usually coming from ImageNet [41] and, mostly, focusing on objects (*e.g.* dog, cat) or scenes (*e.g.* food, snow, beach).

Are, however, generic concepts what we are looking for when delving into our memories? As previous research suggested [10], users search their personal media differently than web media or media coming from one's social circle. However, no previous studies examine such differences rigorously and in real-world scale. To fill in this gap, this paper conducts a real-world comparative study using the unique data of Flickr, which allows search for media in users' personal collection, social circle and multi-billion public media collection on the entire website. With the blessing of Flickr data, we can compare the behaviors in personal search, social search, and web search.

In this paper, we first conduct an in-depth analysis on the Flickr search log data, consisting of 4.3 million queries from more than 133,000 anonymized users. To the best of our knowledge, our study is the first to analyze personal

[1]https://photos.google.com

[2]https://www.flickr.com

[3]In this paper, we use the term *personal media* to refer to both personal photos and videos.

media search using large-scale real-world search log data. By analyzing multiple types of queries and exploring the Flickr personal media repositories, we discover a number of interesting characteristics of personal media search:

1. Personal query sessions are shorter, task-driven and queries are more "visual" than the queries in social or web search.
2. Users are interested in "4W queries" (what, who, where, and when) in their personal media.
3. The majority (about 80%) of personal media have no textual metadata and the percentage with tags is decreasing with time.

Our data analysis also demonstrates a significant gap between personal queries and generic concepts, *i.e.* a gap between a user's information need and what can be retrieved by the system. Traditional text-to-text matching approaches, in which query words are matched against images' metadata, are bound to fail on personal media data as about 80% personal media can only be searched via concepts. Therefore this gap becomes a critical issue hindering the delivery of accurate personal media search.

To bridge this gap, we propose novel approaches based on deep query embedding networks that leverages clickthrough data to learn end-to-end mappings directly from personal queries to the automatic concepts. We propose both feed-forward and Recurrent Neural Network (RNN) [15] architectures to examine the effectiveness of sequential modeling. The proposed model implicitly models the complicated non-linear relations in the visual domain. For example, a user query "birthday party" might not retrieve any results simply because "birthday party" is missing from our concept vocabulary. However, our new approach can translate the query to a set of relevant concepts that exist in the vocabulary, such as "cake", "candle", "kids", etc.

Our experimental results substantiate the efficacy of the proposed models. Our approach consistently outperforms baseline methods in terms of both mean average precision and recall in a test set of over 150 thousand images. Using the proposed models, we see a relative gain of up to 45% over baseline methods in terms of the mean average precision. In summary, our contribution is twofold:

- We conduct an in-depth analysis characterizing differences and similarities between personal media and other types of search.
- We propose a novel visual query embedding framework to bridge the gap between personal queries and concepts. By incorporating deep neural networks, our proposed approach significantly outperforms every baseline in our experiments on a large scale testing set.

We believe our observations may not only be instrumental in guiding future personal media search, but also benefit a variety of related tasks such as personal photo browsing, recommendation, and question answering.

2. RELATED WORK

Although multimedia search has been recognized as an important problem [37], most of existing works focus on text search [44, 10], and very few works focus on personal photo or video search. In their seminal work [10], Dumais *et al.* analyzed personal information retrieval and proposed a system with a unified index for all available information that exploits contextual cues in the search interface. In another early approach [40], the authors proposed a browsing interface tailored for personal media, that uses metadata and basic content analysis. The same ingredients were used by both Begeja *et al.* [2] and Pigeau [39] where they proposed interactive approaches to assist people in organizing and annotating their online personal collections. Utilizing data from Flickr, Maniu *et al.* [34] were interested in discovering behavioral patterns between queries issued in a web multimedia search engine versus a social-sharing site and proposed query-dependent models to improve ranking. Very recently, Bentley *et al.* [3] used the Flickr personal search log to measure the temporal consistency of tags. Although extensive, their analysis and results focused only on temporal patterns of user tags. We are interested in understanding how personal search differs from web search.

The problem of personal media search differs but also can benefit from general web search. Especially, we are interested in analyzing the user behavior, which used to be captured by the clickthrough data in web search. Clickthrough data have been extensively applied for understanding the gap between query and search intent [16, 48, 7, 33, 6, 8]. In one of the earliest approaches, Joachims *et al.* [26] exploited clickthrough data from search logs to learn a ranking SVM for optimizing document search quality. Jiang *et al.* [19] estimated multi-level search satisfaction using Bing search logs and proposed a regression model to predict graded satisfaction. For media search, Jain and Varma [18] used Gaussian Process regression to predict the normalized click count for each result. Yu *et al.* [49] exploited clickthough data to learn multimodal embeddings, whereas O'Hare *et al.* [38] recently utilized interactions like mouse hovering as implicit relevance feedback together with clickthrough data in a learning to rank framework. In this paper, we will use the search log as well as the user clickthrough records, but with the goal of understanding personal media search.

Several innovations were introduced, to modern image search engines, on feature learning, image similarity, and visualization [4, 20, 25, 27, 28, 46]. Most of them can be applied to personal media search. However, the key problem of understanding user queries in personal search can be different from that in web search.

The research of general web search also benefits from the recent progress of modern word embeddings [35]. Very recently, Grbovic *et al.* [14] used web query embeddings together with advertisement click logs to learn query expansion in a distributed system for query to advertisement matching. In a highly related work, Huang *et al.* [17] proposed to learn latent semantic models between queries and documents using clickthrough data. We are interested in learning an visual embedding from the queries to automatically generated *concepts*. We employ concepts as the visual information proxy making the learning space significantly smaller. More importantly, we are able to implicitly take into account the *accuracy* of each concept detector; for example, even if a query exists as one of the concepts, the visual detector for that concept might be weak. We may therefore find other concepts to be equally or even more important for ranking than that query concept itself and improve the ranking of results. We compare against the approach of [17] in Section 6. In some sense, the proposed deep query understanding model might also be the first zero-shot learning approach that uses clickthrough data.

3. DATA

Table 1 summarizes the search log dataset used in our study. We sampled the data from the search logs of the Flickr search from October 2014 to October 2015. From the raw query logs, we extracted the search sessions, and each session contains a anonymized user identifier, query terms, a time stamp, clicked photos/videos along with their ranks and textual metadata in the search result. There are three types of search in the query log: **personal** indicates the queries users issued in searching their own photos; **social** denotes the queries users used to search their friends' photos; **web** indicates the queries searching for anyone's photos on the entire public Flickr collection (billions of photos and videos). We use the Flickr search to approximate web search as the similar search interface and algorithm are used in all web, personal and social search. In total, the dataset contains 4.3 million queries from more than 133,000 users. Flickr is desirable for this study for two reasons: first it is one of the few large-scale websites that offers the personal media search functionality; more importantly, it incorporates different types of search that help distinguish the characteristics of personal media search.

Table 1: Summary of Flickr 4M search log data

Type	Queries	Unique queries	Unique photos
personal	961,826	339,349	820,784
social	560,086	268,183	489,770
web	2,783,525	1,147,386	2,282,881

Both query words and user tags are filtered by a text preprocessing module, which removes the stop words, lemmatizes each word to its root forms and detects a word's POS (Part-of-Speech) tag. Each query is also parsed by a Named Entity Recognizer [11] to extract person, organization and place names. Besides, for each photo and video, we extract 5,000+ image and video concepts by our pre-trained detection models [12, 21, 22]. The concept vocabulary used in our study may be so far one of the largest visual vocabulary. The concepts were trained over tens of millions of images and videos over several big dataset including ImageNet [30], Google Sports 1M [29], YFCC100M [45], DIY [50], etc. In summary, each photo and video in the dataset contains the automatically detected concepts, and, if there is any, user tags and descriptions.

We acknowledge the limitations of the data used in our study. First, the retrieval algorithm used in all types of queries relies on text-to-text matching which might not exactly satisfy the underlying information need about personal media. To reduce this bias, we count queries by the number of *unique* users who issued them, and thus surface more global queries as opposed to user-specific particularities. Second, the Flickr search interface itself has an influence on the analysis results; personal search results are embedded in the generic search interface in a prominent position and therefore intentions can be less clear in some cases.

4. ANALYSIS

4.1 Query Words and User Tags

We start by examining the similarity of the personal, social and web queries. We observed that the query length distributions for different types of queries are similar, in which the queries with less than 4 words account for more than 90% of the total queries. Besides, their POS distributions are also similar in which the top 4 most frequent POS queries are: noun, verb, adj-noun, and noun-noun.

Table 2: Statistics about personal media search.

Query	Personal	Social	Web
Length (w/o stopwords)	1.5	1.8	2.0
Adult Word	0.5%	9.2%	6.8%
Visual	85.3%	60.9%	70.4%

Personal queries are more "visual". A distinguishing characteristic of personal search is that its queries are more "visual". A query is called visual if its information need is about the visual content of the photo or video. For example, "snow", "flower" and "lake district" are visual queries whereas "2014", "NYC", "social media", "Nikon d3200" are not. In order to detect visual queries, we map a POS-tagged query to its closest synset in WordNet, *i.e.* a group of synonym words in WordNet. A query is determined as visual if all of its synsets can be found in the vocabularies of ImageNet [30] and LabelMe [42], the two largest manually curated visual vocabularies of more than 82,000 visual concepts. Our manual analysis on random queries substantiates the coverage of the two visual vocabularies is reasonable.

Table 2 shows 85.3% queries in personal search are visual, which is 15% and 25% higher than that of the web and social queries, respectively. The observation suggests that when users search their own photos, they are more likely to seek visually recognizable content. Since the majority of personal media are not associated with user tags, our finding substantiates the vital role of the automatic concept recognition for personal photos and videos. Besides, we find the personal queries contain significantly less adult words, where the adult words are detected by matching to our dictionary.

The majority of personal media have no user tags and the percentage with tags is decreasing. To estimate the statistics, we randomly sampled a collection of about 200 million personal photos and videos from Flickr. Fig. 1 illustrates the estimated percentage of media with user tags and GPS information, where the x-axis represents the year, and the error bar indicates the 95% confidence interval. On average, 85% personal videos and 77% personal photos do not have any user tag, and the percentage with tags is declining over time probably because of the ever-increasing personal media data. Interestingly, however, the percentage with GPS information is increasing steadily.

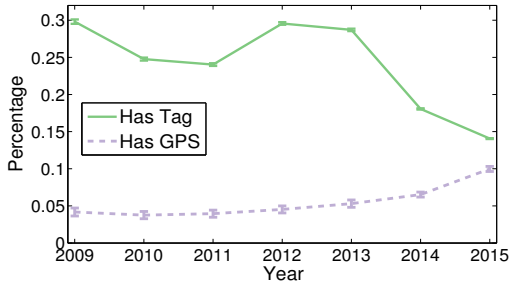


Figure 1: The estimated percentage of personal media data with user tags and GPS information.

Users are interested in "4W queries" in their personal media. To understand the information need of the visual queries, we categories them into "4W" categories, namely, queries searching for *what* (object, thing, action, plant, etc.), *who* (person, animal), *where* (scene, country, city, GPS) and

when (year, month, holiday, date, etc.). To this end, for the synset in the visual query, we traverse the WordNet hierarchy to find its ancestor defined in the top-level synsets in ImageNet and LabelMe. We then manually map the top-level synsets to the 4W categories. For examples, the synset "tree" will be matched to its ancestor synset "plant" which belongs to the *what* category. Besides, the named entities extracted from the query are also used to categorize *where* and *when* queries.

Fig. 2 shows the comparison of 4W distributions. The distribution of frequent user tags is also included for comparison. As we see, the total percentage of 4W queries is much higher in personal queries. The overall distribution is similar across web, social and personal queries but is notably different from the distribution of user tags. The distributional disparity suggests that users tag and search differently. In tagging, the relative percentages of *when* and *where* are higher. This may be because users are accustomed to tag photos by time, place and activities, e.g. "2015 coney island parade" or "fireworks in San Francisco". On the contrary, users search *what* and *who* more often. The distributions in Fig. 2 might influence visual concept vocabulary learning [47, 5]. Besides, the high percentage of *where* in both personal queries and user tags might justify the increased GPS percentage in Fig. 1.

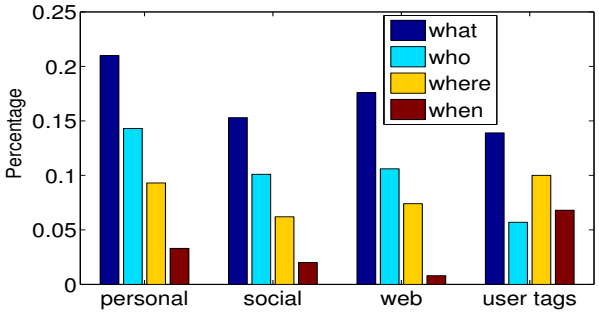


Figure 2: Comparison of 4W categories of personal, social, web queries, and frequent user tags.

4.2 Session Length

This subsection discusses the user click behavior in a search session. Since the search space of the personal media is much smaller than that of the web media, we select queries that return at least 100 results to reduce the analysis bias. As the Flickr search greatly relies on the text-to-text matching, we observed a high matching ratio between the query words and the metadata of the returned photos or videos, which is 88.0% for personal queries and 87.6% for web queries. The ratios indicate that the text-to-text search results are generally relevant for both personal and web queries.

Personal query sessions are shorter. Fig. 3 illustrates the log-log plot of the clicked position in the personal and web query session, where the x-axis denotes the average clicked position in the session, and y-axis represents the percentage. As we see, personal query sessions are considerably shorter than web query sessions in terms of the average clicked position. In addition, the average number of clicks in a personal query session, i.e. 1.57 ± 0.015, is also significantly less than that of the web query session. The statistics suggest that personal search receives fewer clicks and clicked positions in a session are, on average, shallower. This observation may have several implications for personal media

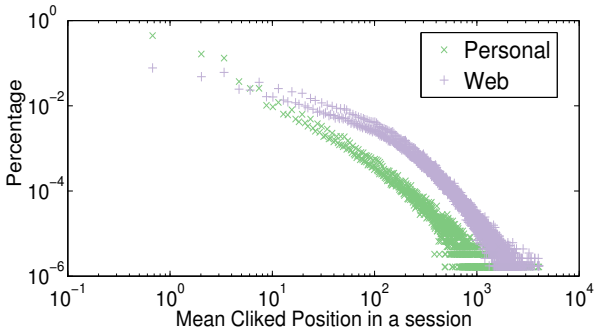


Figure 3: The log-log plot of average clicked position in a session. The median clicked position is 2 for the personal queries and 32 for web queries.

search as it further highlights the importance of the top 2-3 results. Algorithms and interfaces that optimize such results are therefore more desirable.

After inspecting the representative sessions, we hypothesize the underlying reason for the shorter session is that users would ideally search their personal media in a task- or question-driven manner. Our toy survey on 20 active Flickr users substantiates this hypothesis, where the users were asked for queries in an idealized personal media search engine. What we found is the vast majority of queries are either task-driven, e.g. "show me photos with my kid at the playground a couple weeks ago" or in question format "what was the hotel name in our trip to Greece?". Interestingly, for question-driven queries, users seem to be using media search as a mean to recover pieces from their own memories, *i.e.* looking for a specific name, place or date, *e.g.* "who did I have dinner with in WSDM 2016?". The results of this small study suggest that in personal media search, users have clear information need on what to find, either the photos in "show me" queries or answers in question-driven queries, and thus expect shorter search interactions.

4.3 Correlation Analysis

In this subsection we study the correlation between queries, user tags and automatically detected concepts. To reduce mistyped queries and user-specific query terms, we select the 2,000 most frequent query words and user tags ranked by the number of their occurrence in the search log dataset. For concepts, we select a subset of about 4,500 common concept detectors provided by Flickr [12].

We employ three metrics to evaluate the correlation, *i.e.* the Jaccard, WordNet [36] and word2vec embedding [35] similarities. Given two sets of words S_1 and S_2, we calculate the Jaccard coefficient by the size of the intersection divided by the size of the union of the sample sets. For the WordNet and word2vec embedding [35], the correlation is given by:

$$\kappa(S_1, S_2) = \frac{1}{2|S_1|} \sum_{w_i \in S_1} \max_{w_j \in S_2} \kappa(w_i, w_j), \qquad (1)$$

where $\kappa(w_i, w_j)$ represents the generic similarity algorithm, *i.e.* the WordNet shortest path similarity [31] or the cosine word2ve similarity in the pre-trained embedding on Google News. The final correlation is equal to $\kappa(S_1, S_2) + \kappa(S_2, S_1)$.

The above metrics capture the correlation from different perspectives. Jaccard measures the proportion of exactly matched words. WordNet captures the word similarity in terms of the length of the shortest path that connects the word senses in the is-a taxonomy, and therefore is good

at capturing synonyms and subsumption relations between nouns. On the other hand, word2vec examines the latent semantic relatedness of two words in a low-dimensional embedding space. This metric is good at capturing the words that co-occur frequently in similar contexts.

We present the correlations in Table 3. As we see, the metrics seems to follow a similar pattern in which the correlation between web & personal queries, and personal queries & user tags is higher. On the contrary, the correlation between automatically detected concepts is about 20% lower. The results suggest that there is a significant gap between personal queries and automatically detected concepts, *i.e.* a gap between a user's information need and what can be detected by the system. Given the fact that about 80% personal media can only be searched via such concepts, such a gap harms performance in personal search. Our experimental results in Section 6 substantiate this argument.

Table 3: Correlation between personal query words, user tags and concepts.

Type	Jaccard	WordNet	Word2vec
personal & web	0.326	0.701	0.696
personal & tags	0.336	0.687	0.702
personal & concepts	0.197	0.528	0.542
tags & concepts	0.144	0.443	0.490

4.4 Summary of Findings

In this section, we analyzed the click logs and present similarities and differences between personal and other types of media search. We found that personal media queries are more "visual" and have a higher percentage of correspondence to the "4W" categories, *i.e.* what, who, when and where. We estimated that about 80% personal photos and videos do not have any user-generated tags, and this percentage increases over time. After analyzing clickthrough data in sessions, we found personal media search to receive significantly fewer clicks and the average clicked position is shallower than web media search. We hypothesize that personal queries are usually task- or question-driven over seen photos or videos, as opposed to the exploratory nature of a large percentage of web media searches. Users have clear information need on what to find, and thus expect shorter search interactions. Finally, the correlation analysis between personal queries and automatically detected concepts indicates a significant gap between user information needs and what can be retrieved by the current system.

5. DEEP QUERY UNDERSTANDING

To bridge the gap between personal query words and concepts discussed in Section 4, in this section, we introduce a new method, named Visual Query Embedding (VQE), to improve the personal media search.

5.1 Problem Formulation

A concept corresponds to a visual recognition model that estimates the probability of observing the concept in the image or video content. There are two major differences between the concepts in personal media and the words in text documents. First, the concept vocabulary is much smaller than the word vocabulary, limited by the number of objects, scenes or actions that can be accurately detected in the content of photos or videos. Scaling the number of concepts is nontrivial, as training detectors requires considerable amount of labeled data which are expensive to acquire [32]. Second, the accuracy of the automatically detected concepts is limited: the detected concepts may not actually be present whereas concepts not detected may well appear in the content of personal media.

Due to such differences, there is a significant gap between personal query words and concepts, *i.e.* a gap between a user's information need and what can be retrieved by the system. To address this issue, we propose to learn Visual Query Embedding (VQE) models that directly map user query words to the related visual concepts. We propose to address this problem through a novel perspective where end-to-end embeddings are learned leveraging visually relevant concepts discovered in the clickthrough data. Following [17], we assume a query to be relevant, at least partially, to clicked personal media data in that session. Our intuition is that for the same query, concepts frequently occurring in the clicked photos are more likely to be relevant. For example, if many users clicks photos containing the concept "candles" for the query "birthday party", then "candles" is a concept that is probably related to "birthday party". Table 4 shows some representative examples discovered form the search logs.

Table 4: Examples of user queries and visually relevant concepts.

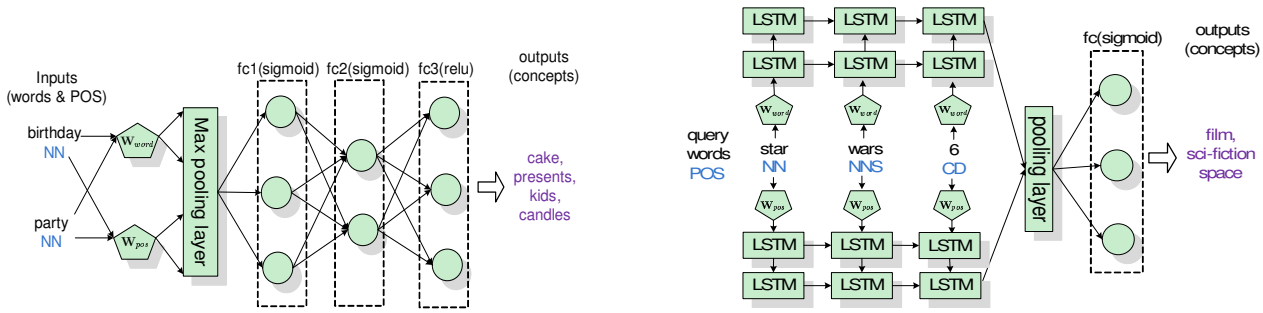
User queries	Related Visual Concepts
jaguar →	sports car, road
playa →	coast, ocean
bluebell →	flower, purple
tiger →	carnivore, big cat, tiger
andromeda →	empty, dreamlike, fire, bonfire
zoo →	people, animal, primate, dog, monkey

We are interested in learning an end-to-end visual query embedding function from the user query words to the relevant visual concepts discovered in the clickthrough data. Formally, let $Q = q_1, \cdots, q_n$ denote a query of n words, where $Q \in \mathbb{Z}^n$ and each q_i represents an integer index in the query word vocabulary. Define a function $\phi : \mathbb{Z}^n \rightarrow \mathbb{R}^m$, where $\mathbb{R}^m$ is a vector over the concept vocabulary of m concepts. Denote $\mathbf{y}_k$ as the relevant concepts extracted from the search log for the query Q_k. Based on the above definitions, we can summarize the visual query embedding problem as a supervised learning problem: $\phi = \arg\min_\phi \sum_k \ell(\phi(Q_k), \mathbf{y}_k)$, where ℓ is the loss function.

In the online search phase, given a user query Q_k, we use ϕ to map it to a vector of relevant concepts, and apply retrieval algorithms to obtain the relevant personal media. In this paper, we employ the vector space (cosine) retrieval model for simplicity, and refer readers to [23] for an analysis on the impact of retrieval algorithms.

5.2 Visual Query Embedding Models

This subsection discusses two deep models for learning the visual query embedding. First of all, we introduce a method to extract visually relevant concepts in a search session. The personal photos or videos are automatically tagged by m concepts in $\mathbb{V}$. Let $\mathbf{d} \in \mathbb{R}^m$ represent the raw detection scores, each dimension of which corresponds to the probability of detecting a concept. For a photo or video, $\mathbf{d}$ is usually a dense vector. In other words, a photo contains almost every concept in the vocabulary with a non-zero detection score. We found learning ϕ based on the dense raw score representation not only leads to worse results but



(a) Max-Pooled MLP (b) Two-channel RNN

Figure 4: Deep Visual Query Embedding Models.

also becomes infeasible for large-scale learning. To address this issue, we incorporate the concept adjustment method in [24], and represent personal media data by the adjusted concept vector $\mathbf{v} \in \mathbb{R}^m$, given by:

$$\arg\min_{\mathbf{v}\in[0,1]^m} \frac{1}{2}\|\mathbf{v}-\mathbf{d}\|_2^2 + \alpha\|\mathbf{v}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{v} \leq \mathbf{0}, \tag{2}$$

where α is the parameter that controls the sparsity. For simplicity, we follow the algorithm in [24], use the l_1-norm, and set $\mathbf{A}$ to be the zero matrix as most of the concepts in our experiments are independent.

In the kth session, let C_k^+ represent a set of adjusted concept vectors in the clicked personal media. We define the ground truth vector as the mean of the clicked concept vectors, i.e. $\mathbf{y}_k = 1/|C_k^+|\sum_{\mathbf{v}_i\in C_k^+}\mathbf{v}_i$. Note that the click-through data are very noisy [43], containing many queries and clicks made by errors. We found the quality of the training set to greatly affect the accuracy the learned visual query embedding. To reduce noise, we select queries issued by at least 3 users, only consider clicks on the top 30 retrieved results, and the concepts that occur in at least two clicked photos in a session. Within a session, for each concept we compute the mutual information in the set of clicked media and a set of randomly sampled non-clicked media. A concept with lower mutual information means it occurs, indiscriminately, in both clicked and non-clicked sets, and thus is likely to be a background concept such as "outdoor" and "people". For training, we zero the background concepts with small mutual information in the ground-truth vector $\mathbf{y}$.

Given a training set of N sessions, let $\hat{\mathbf{y}}_i$ represent the embedding output after the softmax activation function, i.e. $\hat{\mathbf{y}}_k = \text{softmax}(\phi(Q_k))$, the embedding is learned by minimizing the *cross-entropy loss* function:

$$\arg\min_{\phi} \sum_{i=1}^{N} \ell(\hat{\mathbf{y}}_i, \mathbf{y}_i) = -\sum_{i=1}^{N}\sum_{j=1}^{m} \mathbb{1}(y_{ij}>0)\log\hat{y}_{ij} + \mathbb{1}(y_{ij}=0)\log(1-\hat{y}_{ij}) \tag{3}$$

where $\mathbb{1}(\cdot)$ is an indicator function equaling 1 when its argument is true, and 0 otherwise. Eq. (3) is also known as softmax cross-entropy loss. In the rest of the section, we will discuss two deep neural networks to learn the model.

5.2.1 Max-Pooled MLP

Our first model is the max-pooled Multi-Layer Perceptron (MLP), with architecture depicted in Fig. 4(a). It takes query words and their Part-of-Speech (POS) tags as input, and outputs the predicted concept vector. The model consists of three types of layers: an embedding layer which maps a word or a POS tag to a low-dimensional vector; a max pooling layer that computes the element-wise maximum for the input vectors; a number of fully connected layer (fc) for nonlinear transformation. Due to the disjoint vocabulary space, we learn separate embeddings $\mathbf{W}_{word}$ for query words and $\mathbf{W}_{pos}$ for POS tags. Denote q_i as the ith word and p_i as its POS tag in the query Q, the model with l layer is calculated from:

$$\begin{aligned} \mathbf{a}_1 &= \max_{q_i,p_i\in Q}(\mathbf{W}_{word}(q_i), \mathbf{W}_{pos}(p_i)) \\ \mathbf{a}_i &= \sigma(\mathbf{W}_i\mathbf{a}_{i-1} + \mathbf{b}_i), \\ \phi(Q) &= \text{relu}(\mathbf{W}_l\mathbf{a}_{l-1} + \mathbf{b}_l) \end{aligned} \tag{4}$$

where $\sigma(x) = (1+e^{-x})^{-1}$ is the sigmoid activation function in the hidden layers, and relu is the rectified linear unit in the last layer. $\mathbf{W}_i$ and $\mathbf{b_i}$ represent the weight matrix and the bias term vector in the ith layer; $\mathbf{a}_i$ is the activation of the ith layer, and $\phi(Q)$ is the predicted concept vector.

5.2.2 Two-channel RNN

The word sequence in a query is totally discarded by the max-pooling layer in the previous model. To incorporate the sequence information, we propose a two-channel RNN model. As illustrated in Fig. 4(b), the embedding vectors of the word and POS tags are fed into a two layer LSTM units one by one, via two channels: $[q_1,\cdots,q_n,\$]$ and $[p_1,\cdots,p_n,\$]$, where \$ is a special token that marks the end of a sequence. LSTM units are used to reduce the vanishing gradients and exploding gradients problem [15]. More precisely, we use the LSTM unit with dropout implementation described in [51]. LSTM updates for time step t, given a word or pos embedding vector as the inputs $\mathbf{x}_t$:

$$\begin{aligned} \mathbf{i}_t &= \sigma(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{c_i}\mathbf{c}_{t-1} + \mathbf{b}_i) \\ \mathbf{f}_t &= \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \\ \mathbf{c}_t &= \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t\tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{c}_{h-1} + \mathbf{b}_c) \\ \mathbf{o}_t &= \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o) \\ \mathbf{h}_t &= \mathbf{o}_t\tanh(\mathbf{c}_t) \end{aligned} \tag{5}$$

where $\mathbf{i}, \mathbf{f}, \mathbf{o}$ and $\mathbf{c}$ are respectively the input gate, forget gate, output gate and memory cell activation vectors. All of which are the same size of the hidden vector $\mathbf{h}$.

The LSTM hidden states $\mathbf{h}_t$ from the input sequences are fed into an average pooling layer, as shown in Fig. 4(b). The pooled hidden states are then fed to a set of fully connected layers similar to those in Eq. (4). The final predicted concept vector $\phi(Q)$ is derived from the output of the final fully connected layer.

We incorporate the POS tags in our models for two considerations: first, we found, though adding POS tags would slow down the convergence, in some cases, it helps to find better local minima. The second reason is for generalizability. The proposed models can trivially degenerate to the models without POS tags when tags are less informative.

6. EXPERIMENTS

6.1 Experimental Setup

Dataset and evaluation: We conduct our experiments on the Flickr personal search log data. We select personal queries that were issued by at least 3 users, and divide them into a training and a test set according to their issued time. In total, the training set contains about 20,600 personal queries from 3,978 users, while the test set contains 2,443 queries from 1620 users over about 148,000 personal photos. Given a personal query and a photo collection from a user, our goal is to boost the rank for the user clicked photos. We discard all user generated textual metadata that may exist in the user photos in our experiments, and only assume that each photo is tagged with 1,720 automatically detected concepts sampled from the Flickr concept bank [12].

We evaluate performance using two metrics: the non-interpolated mean Average Precision (mAP) of the retrieved ranked list and the concept recall of the top predicted concepts denoted as CR@n. Let $\mathbf{t}$ represents the predicted concepts $\phi(Q)$ after the top-n thresholding, *i.e.* all elements except for the top n elements in $\phi(Q)$ are set to 0, we have:

$$\text{CR@n}(\mathbf{t}, \mathbf{y}) = \frac{1}{n} \sum_{j=1}^{m} \mathbb{1}(y_i t_i > 0), \tag{6}$$

where $\mathbf{y}$ is the ground-truth concept vector extracted in the search session, and $\mathbb{1}$ is the indicator function. Note the two metrics measures different aspects of the search results. mAP evaluates the quality of the clicked photos ranked in the search results, whereas CR@n measures the relevance between the top-n predicted concepts and the true concepts in the clicked photos. A relevant concept may not always lead to a good ranked list as it might be less discriminative, e.g. the relevant concept "carnivore" to the query "tiger". On the other hand, discriminative concepts leading to better mAP may not always be relevant. Therefore, both metrics are useful in understanding the performance of a method.

Compared Methods: We refer to the two Visual Query Embedding (VQE) models discussed in Section 5, as VQE (MaxMLP) and VQE (RNN). To demonstrate their performance, we compare them against the following common zero-shot learning and word embedding approaches: **Exact Match** [23] is a plain mapping by matching the exact query words to the concept names. Specifically, it produces a query vector of the same size with the concept vocabulary, each dimension of which represents the similarity between the query and the corresponding concept. The generated query vector is then used to search relevant personal photos. Likewise, **WordNet** computes similarities between query vectors and concepts using WordNet path similarity [36] which is equal to the shortest path in the WordNet taxonomy between the query and the concept name [36]. **SkipGram** [35] learns an embedding space over a large corpus of text documents. In our experiments, the pretrained embedding on GoogleNews is used to compute the query vector. **Semantic DNN** is inspired by the deep semantic structured model of [17], where the authors proposed to learn a low-dimensional embedding space form the query words to the words in the clicked text documents by multilayer neural networks. In our problem the vocabularies of query and concept are different, and as a result, we add a layer on top of the last layer of the DNN model in [17] to obtain the predicted concept vector. As in [17], the cosine loss function is used to train the model. Note that only the VQE models and the semantic DNN model use the clickthrough training data.

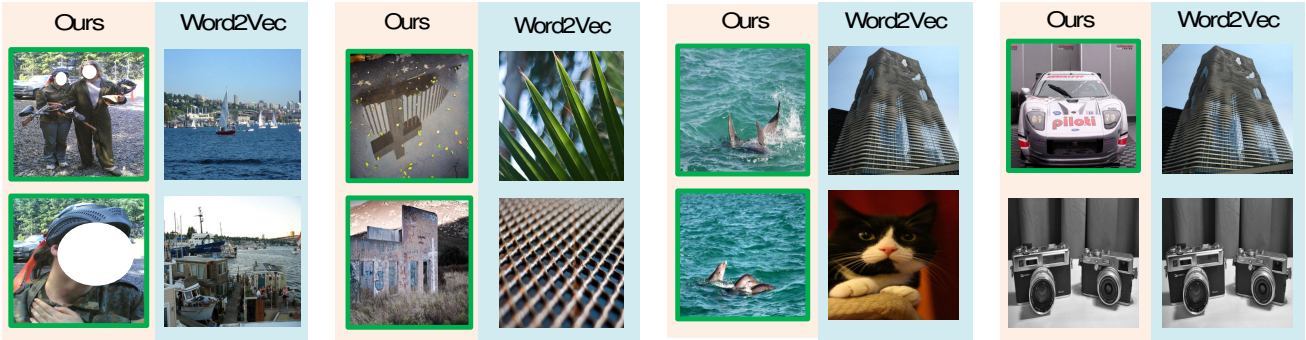
Implementation Details: We implement the proposed VQE models in TensorFlow [1]. The model are trained over mini-batches of 32 samples. The word and POS embeddings are set to 300 dimensional vector and are learned jointly by minimizing the loss in Eq. (3). The standard gradient decent algorithm is used to train the MLP models, and the adaptive subgradient (Adagrad) [9] algorithm is used to train the RNN models for faster convergence. The start learning rate is set to 0.1 and is annealed by a staircase exponential decay function with a decay rate of 0.96. A dropout layer is applied in training the RNN networks which discards 0.5% of the input data. Each model is trained at most 7200 epochs (no more than 24 hours).

6.2 Baseline Comparisons

We first compare the proposed methods with the baseline methods in Table 5. As we see, the proposed VQE MaxMLP significantly outperforms other baseline methods. Specifically, it improves the mAP of SkipGram by about relative 45%. We inspected the search results and found that MaxMLP can capture more visually relevant concepts for personal media queries. Fig. 5 shows representative examples of the top search results for MaxMLP and SkipGram models, where the photos in the green border are the user clicked photos in the search session. As shown in Fig. 5(a), MaxMLP retrieves more accurate personal photos. This is because it maps the user query "paint ball" to visually relevant concepts "solider" and "fatigues", as opposed to the concepts "archery" and "skateboarding" produced by SkipGram. In addition, we found MaxMLP model can find relevant concepts for "who" and "where" quires (see Fig. 2), the two major categories in personal queries. For example, as shown in Fig. 5(c), the MaxMLP model maps the user query "key west", *i.e.* a island city, to the concepts "water" and "water sports", whereas SkipGram fails to find any relevant concept. Besides, experimental results also show the domain difference between learning embedding on clickthrough data versus learning embedding on text corpora like Google News.

Although the proposed method shows promising results. We admit that it is still significantly worse than traditional text-to-text search over the photos or videos with rich user-generated metadata. We believe the problem is novel, challenging, and needs further research [23]. We found the lack of common sense often results in inaccurate mappings in the VQE (MaxMLP) model. For example, the user query "bus" is mapped to "tramline" by the VQE model even though there exists a "bus" concept in the vocabulary. This problem may be addressed by either incorporating prior knowledge in training or by increasing the size of the training data. Besides, the worse performance of Semantic DNN model might stem from the less appropriate loss function. See Section 6.3 for more discussions.

The proposed VQE (RNN) model yields better CR@1 and



(a) Paintball (b) Cross (c) Key west (d) Ford

Figure 5: Examples of top search results of personal photos. The left ranked list indicates our results and the right list is from the SkipGram (word2vec). The user query is listed in the subtitle, and the photos in the green border are the user clicked photos.

CR@3 but worse mAP than the baseline SkipGram method, suggesting that the RNN model can find relevant but less discriminative concepts. We found two reasons explaining the worse performance of VQE (RNN) when compared to the VQE (MaxMLP) model: first the worse results suggest the word sequence in personal queries is less informative. It is acknowledged that the sequence of text query words plays a less important role in the bag-of-word or unigram language retrieval model [52]. Our experimental results suggest this may still hold in personal media search. Second, the RNN model converges much slower than the MaxMLP model. When we stopped the training for the RNN model after 24 hours, its performance is still worse than that of the MaxMLP model.

Table 5: Comparison to baseline methods.

Method	mAP	CR@1	CR@3	CR@5
Exact Match [23]	0.231	0.209	0.086	0.067
WordNet [36]	0.269	0.298	0.195	0.161
SkipGram [35]	0.271	0.286	0.194	0.173
Semantic DNN [17]	0.120	0.010	0.018	0.018
VQE (RNN)	0.235	0.377	0.238	0.167
VQE (MaxMLP)	**0.390**	**0.524**	**0.374**	**0.289**

6.3 Model Parameters

In this section, we study the impact of parameters in the proposed VQE models. First we empirically compare neural network structures. Table 6 lists different neural network structures of VQE models, where embedding layers and polling layers are omitted to save space. The detailed model for MaxMLP and RNN model can be found in Eq. (4), Eq. (5), and Fig. 4. For example, the third row MaxMLP4 represents a 4-layer network containing an embedding layer, a pooling layer, a fully connected layer fc_1, transforming a 300d max-polled vector to a hidden layer of 300d by the sigmoid function, and an output layer fc_2, transforming the 300d hidden vector to a output vector of 1720d by the rectified linear unit. The fifth row MeanMLP4 represents the same network as MaxMLP5 except that it employs the mean instead of the max polling layer .

We observed two trends in Table 6. First the performance increases as models get deeper. This observation suggests the visual query embedding for personal media can be highly nonlinear, and deeper models may better capture the underlying relation between user query words and relevant concepts. For example, the 5-layer MaxMLP5 achieves better mAP than the 4-layer MaxMLP4. However, in fact, MaxMLP5 has fewer parameters than MaxMLP4. Second, we found the max polling in the MaxMLP model leads to not only faster convergence but also more accurate search results. For example, MaxMLP5 outperforms MeanMLP5 suggesting the efficacy of the max-polling layer.

Table 6: Comparison of network structures.

Model	Network Structure	mAP	CR@3
MaxMLP3	fc_1: relu(300 $\rightarrow$ 1720)	0.225	0.314
MaxMLP4	fc_1: sigmoid(300 $\rightarrow$ 300) fc_2: relu(300 $\rightarrow$ 1720)	0.367	0.301
MaxMLP5	fc_1: sigmoid(300 $\rightarrow$200) fc_2: sigmoid(200 $\rightarrow$200) fc_3: relu(200 $\rightarrow$1720)	**0.390**	**0.374**
MeanMLP5	Same as above.	0.249	0.202
RNN3	$lstm_1$ lstm:(300 $\rightarrow$ 200) fc_1: relu(200 $\rightarrow$ 1720)	0.124	0.025
RNN6	$lstm_1$ lstm:(300 $\rightarrow$ 200) $lstm_2$ lstm:(200 $\rightarrow$ 200) fc_1: sigmoid(200 $\rightarrow$ 200) fc_2: sigmoid(200 $\rightarrow$ 200) fc_3: relu(200 $\rightarrow$ 1720)	0.235	0.238

The loss function is an important component in neural network training. The softmax cross-entropy loss discussed in Eq. (3) represents a type of loss that jointly models concepts as a sparse vector due to the softmax transformation. Alternatively, we can use the cross-entropy loss, which ignores the sparse constraint, or the cosine loss, which measures the distance between queries and concepts seen as dense vectors. Our goal is to find which type of loss is suitable for VQE learning. Table 7 lists the mAP performance. As we see, the cosine loss yields the worst results suggesting treating concepts as dense vectors in the high dimensional space is less appropriate in our problem. This may explain the worse performance of Semantic DNN in Table 5. Besides, the comparison between the cross-entropy and the softmax cross-entropy suggests jointly modeling concepts as a sparse representation is helpful.

Table 7: mAP for different loss functions.

Loss Function	MLP	RNN
Softmax cross-entropy	0.390	0.235
Cross-entropy	0.187	0.145
Cosine distance	0.124	0.130

7. CONCLUSIONS

In this paper we investigated personal media search using clickthrough data on a large-scale, real-world set. We analyzed different types of search sessions mined from Flickr search logs and discovered a number of interesting attributes of personal media search. We found personal queries to be more visual and task- or question-oriented, aiming visual semantics in a small collection of media, the majority of which have no tags or descriptions.

We further found that automatically generated concepts, one of the very few options for searching in the absence of other textual metadata, cannot properly capture the user intent in personal search. Inspired by our findings, we proposed novel models for learning visual query embeddings between user queries and concepts and achieved high gains in search performance over existing baselines methods.

To our knowledge, this paper is the first to study the nature of personal search using a large amount of real-world data and gives insightful observations that enable us to learn novel deep visual query embeddings that can improve search performance. However, the proposed model is still significantly worse than the traditional text search over the photos or videos with rich user-generated metadata. We believe the problem is novel, challenging, and needs further research. Besides, the insights from our analysis further show that question-based search is a very important aspect of personal search. We believe that further research on question answering over personal media is needed, and we expect this to be a promising and fruitful direction.

8. ACKNOWLEDGMENTS

This work was partially supported by Yahoo InMind project. The authors would like to thank the members of the Flickr Computer Vision and Machine Learning group for their kind support and meaningful discussions.

References

[1] M. Abadi et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] L. Begeja, E. Zavesky, Z. Liu, D. Gibbon, R. Gopalan, and B. Shahraray. Vidcat: an image and video analysis service for personal media management. In *IS&T/SPIE Electronic Imaging*, 2013.

[3] F. Bentley, D. A. S. Joseph Jofish Kaye, and J. A. Guerra-Gomez. The 32 days of christmas: Understanding temporal intent in image search queries. In *CHI*, 2016.

[4] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010.

[5] J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, 2014.

[6] H. Cheng and E. Cantú-Paz. Personalized click prediction in sponsored search. In *WSDM*, 2010.

[7] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR*, 2007.

[8] O. Dan, V. Parikh, and B. D. Davison. Improving ip geolocation using query logs. In *WSDM*, 2016.

[9] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011.

[10] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR*, 2003.

[11] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[12] P. Garrigues, S. Farfade, H. Izadinia, K. Boakye, and Y. Kalantidis. Tag prediction at flickr: a view from the darkroom. *arXiv preprint arXiv:1612.01922*, 2016.

[13] Google. Google photos: One year, 200 million users, and a whole lot of selfies. https://googleblog.blogspot.com/2016/05/google-photos-one-year-200-million.html, 2016.

[14] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, R. Baeza-Yates, A. Feng, E. Ordentlich, L. Yang, and G. Owens. Scalable semantic matching of queries to ads in sponsored search advertising. In *SIGIR*, 2016.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li. Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In *MM*, 2013.

[17] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, 2013.

[18] V. Jain and M. Varma. Learning to re-rank: query-dependent image re-ranking using click data. In *WWW*, 2011.

[19] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM*, 2015.

[20] L. Jiang. Web-scale multimedia search for internet video content. In *WWW*, 2016.

[21] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014.

[22] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[23] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ICMR*, 2015.

[24] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. Hauptmann. Fast and accurate content-based semantic search in 100m internet videos. *MM*, 2015.

[25] Y. Jing, H. Rowley, J. Wang, D. Tsai, C. Rosenberg, and M. Covell. Google image swirl: A large-scale content-based image visualization system. In *WWW*, 2012.

[26] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[27] Y. Kalantidis, L. Kennedy, H. Nguyen, C. Mellina, and D. A. Shamma. Loh and behold: Web-scale visual search, recommendation and clustering using locally optimized hashing. *arXiv preprint arXiv:1604.06480*, 2016.

[28] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias. Viral: Visual image retrieval and localization. *Multimedia Tools and Applications*, 51(2):555–592, 2011.

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[31] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

[32] J. Liang, L. Jiang, D. Meng, and A. Hauptmann. Learning to detect concepts from webly-labeled video data. In *IJCAI*, 2016.

[33] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query logs. In *WSDM*, 2011.

[34] S. Maniu, N. O'Hare, L. M. Aiello, L. Chiarandini, and A. Jaimes. Search behaviour on photo sharing platforms. In *ICME*, 2013.

[35] T. Mikolov and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[36] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

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

[38] N. O'Hare, P. de Juan, R. Schifanella, Y. He, D. Yin, and Y. Chang. Leveraging user interaction signals for web image search. In *SIGIR*, 2016.

[39] A. Pigeau. Life gallery: event detection in a personal media collection. *MTAP*, pages 1–22, 2016.

[40] J. C. Platt, M. Czerwinski, and B. A. Field. Phototoc: Automatic clustering for browsing personal photographs. In *ICSP*, 2003.

[41] 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. *IJCV*, 115(3):211–252, 2015.

[42] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[43] A. Singla and R. W. White. Sampling high-quality clicks from noisy click data. In *WWW*, 2010.

[44] Y. Song, H. Wang, and X. He. Adapting deep ranknet for personalized search. In *WSDM*, 2014.

[45] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[46] G. Tolias, Y. Kalantidis, Y. Avrithis, and S. Kollias. Towards large-scale geometry indexing by feature selection. *Computer Vision and Image Understanding*, 120:31–45, 2014.

[47] B. Varadarajan, G. Toderici, S. Vijayanarasimhan, and A. Natsev. Efficient large scale video classification. *arXiv preprint arXiv:1505.06250*, 2015.

[48] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. In *CIKM*, 2004.

[49] J. Yu, Y. Rui, and B. Chen. Exploiting click constraints and multi-view features for image re-ranking. *IEEE Transactions on Multimedia*, 16(1):159–168, 2014.

[50] S.-I. Yu, L. Jiang, and A. Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *MM*, 2014.

[51] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[52] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, 2001.