Delving Deep into Personal Photo and Video Search

Lu Jiang, Yannis Kalantidis, Liangliang Cao, Sachin Farfade, Jiliang Tang, Alexander G. Hauptmann
Yahoo, CMU and MSU
Outline

- Introduction
- Flickr Data
- Statistical Characteristics
- Deep Query Understanding
- Experiments
- Conclusions
Outline

- Introduction
- Flickr Data
- Statistical Characteristics
- Deep Query Understanding
- Experiments
- Conclusions
Introduction

"zootopia"

13.9 petabytes personal media were uploaded to Google photo by 200M users in just one year.

More than 80% personal media do not have user tags.

reality
Introduction

• Personal media: personal photos and videos.

• Personal media search is a novel and challenging problem:
  – what are the differences when users search their own photos/videos versus the ones on the web?
  – Can we use our findings to improve personal media search?

• Conduct our research on large-scale real-world Flickr search logs.
Outline

- Introduction
- Flickr Data
- Statistical Characteristics
- Deep Query Understanding
- Experiments
- Conclusions
Data

Large-scale real-world search log data on Flickr:

<table>
<thead>
<tr>
<th>Type</th>
<th>Queries</th>
<th>Unique queries</th>
<th>Unique photos</th>
</tr>
</thead>
<tbody>
<tr>
<td>personal</td>
<td>961,826</td>
<td>339,349</td>
<td>820,784</td>
</tr>
<tr>
<td>social</td>
<td>560,086</td>
<td>268,183</td>
<td>489,770</td>
</tr>
<tr>
<td>web</td>
<td>2,783,525</td>
<td>1,147,386</td>
<td>2,282,881</td>
</tr>
</tbody>
</table>

- **personal**: queries searching their own photos
- **social**: searching their friends’ photos
- **web**: searching anyone’s photos on the entire public Flickr
**Data**

**Concepts** are automatically detected objects, people, scenes, actions, etc. from images or videos.

The precision of recall detected concepts are **limited**. As most photos have no metadata, the concept is one of the few options. We extract **5000+ concepts** from the image and video on the data.
Outline

- Introduction
- Flickr Data
- Statistical Characteristics
- Deep Query Understanding
- Experiments
- Conclusions
Observation I

- personal queries are more “visual”.

<table>
<thead>
<tr>
<th>Query</th>
<th>Personal</th>
<th>Social</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>85.3%</td>
<td>60.9%</td>
<td>70.4%</td>
</tr>
</tbody>
</table>

**visual**: snow, flower, lake

**non-visual**: 2014, NYC, social media

Visual Concept Vocabularies

query words → WordNet synsets → LabelMe
Observation II

• The majority of personal media have no user tags and the percentage with tags is decreasing.

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

• Users are interested in “4W queries” in their personal media.
  – what (object, thing, action, plant, etc.)
  – who (person, animal)
  – where (scene, country, city, GPS)
  – when (year, month, holiday, date, etc.)

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

• Personal search sessions are shorter. The median clicked position is 2.

• Getting the top2 personal photos correct is very important.
Observation V

- A big **gap** between **personal queries** and automatically **detected concepts**.
  - Users might search millions of topics, but the system can detect only a few thousand concepts.
Outline

- Introduction
- Flickr Data
- Statistical Characteristics
- Deep Query Understanding
- Experiments
- Conclusions
Problem

• To reduce the gap, we can:
  – Increase the #concepts $\rightarrow$ non-trivial (more labeled data) 😞
  – Better understand the query $\rightarrow$ **focus** of this paper

• **query understanding**: how to map out-of-vocabulary query words to the concepts in our vocabulary?

**user query**

Making a sandwich

**generated query**

food, bread, cheese, kitchen, cooking, room, lunch, dinner;

all concepts are in our vocabulary
Query understanding is a challenging problem. Existing methods include:

- **Exact word matching**
- **WordNet Similarity** [Miller, 1995]: structural depths in WordNet taxonomy.
- **Word embedding mapping** [Mikolov et al., 2013]: word distance in a learned embedding space in Wikipedia by the skip-gram model (word2vec).


T. Mikolov and J. Dean. Distributed representations of words and phrases and their compositionality. 2013.
Deep Visual Query Embedding

• Our solution: learning deep visual query embedding mined from the Flickr search logs.

User query: making a sandwich

Concepts in clicked photos: bread, vegetable, kitchen, cooking, ham

• Training data: mined from Flickr search logs.

<table>
<thead>
<tr>
<th>User queries</th>
<th>Related Visual Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>jaguar →</td>
<td>sports car, road</td>
</tr>
<tr>
<td>playa →</td>
<td>coast, ocean</td>
</tr>
<tr>
<td>bluebell →</td>
<td>flower, purple</td>
</tr>
<tr>
<td>tiger →</td>
<td>carnivore, big cat, tiger</td>
</tr>
<tr>
<td>andromeda →</td>
<td>empty, dreamlike, fire, bonfire</td>
</tr>
<tr>
<td>zoo →</td>
<td>people, animal, primate, dog, monkey</td>
</tr>
</tbody>
</table>
Max-Pooled MLP
Two-channel RNN
Outline

- Introduction
- Flickr Data
- Statistical Characteristics
- Deep Query Understanding
- Experiments
- Conclusions
Experimental Setups

• Task: search 148,000 personal photos by concepts. We discard all textual metadata.

• Goal: rank the user clicked photos closer to the top.
  – Training: 20,600 queries from 3,978 users
  – Test on 2,443 queries from 1,620 users.

• Evaluated by the mean average precision (mAP) and the concept recall at k (CR@k).
  • mAP → how well the clicked photos are ranked in the results.
  – CR@k → how accurate are the top-k predicted concepts.
Results

Word2vec embedding on GoogleNews

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>R@1</th>
<th>R@3</th>
<th>R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Match [19]</td>
<td>0.231</td>
<td>0.209</td>
<td>0.086</td>
<td>0.067</td>
</tr>
<tr>
<td>WordNet [28]</td>
<td>0.269</td>
<td>0.298</td>
<td>0.195</td>
<td>0.161</td>
</tr>
<tr>
<td>SkipGram [27]</td>
<td>0.271</td>
<td>0.286</td>
<td>0.194</td>
<td>0.173</td>
</tr>
<tr>
<td>Semantic DNN [15]</td>
<td>0.120</td>
<td>0.010</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td>VQE (RNN)</td>
<td>0.235</td>
<td>0.377</td>
<td>0.238</td>
<td>0.167</td>
</tr>
<tr>
<td>VQE (MaxMLP)</td>
<td>0.390</td>
<td>0.524</td>
<td>0.374</td>
<td>0.289</td>
</tr>
</tbody>
</table>

- mAP is pretty low → challenging problem.
- Learning the deep embedding over the search logs helps.
- Considering the word sequence might not help.
- RNN model converges much slower and thus gets worse performance.
Examples of top 2 results

(a) Paintball

(c) Key west
Empirical Observations

- **deeper** models are better.
- **max pooling** is better than average pooling.
- **softmax loss** yields the best results \(\rightarrow\) probably because concepts in the clicked photos are sparse.
Outline

- Introduction
- Flickr Data
- Statistical Characteristics
- Deep Query Understanding
- Experiments
- Conclusions
Take home messages

• Personal media search:
  ▪ Personal query sessions are shorter, and queries are more “visual”.
  ▪ Users are interested in “4W queries”.
  ▪ 80% of personal media have no textual metadata and the percentage is decreasing.

• Utilizing deep learning for query understanding is promising in improving personal media search.

• We believe personal media search is a novel and challenging problem, which needs further research.
Thank You.

Questions to: lujiang@cs.cmu.edu

Check out our application. Search MemoryQA on YouTube.