Known-Item Search in Image Datasets Using Automatically Detected Keywords

Tomáš Souček
Faculty of Mathematics and Physics, Charles University

Introduction

Our goal was to design and evaluate a **keyword retrieval model for known-item search** (KIS) in image collections. Specifically, we have done:

- ▷ Selected a large-scale image dataset with help of machine learning.
- ▷ Trained a deep neural network to predict 1150 image classes.
- ▷ Designed query interface for fast image retrieval.
- ▷ Estimated the model's performance in an interactive setting by designing several types of artificial users.

Dataset Selection

We considered 6642 image classes, each with more than 1000 examples from ImageNet database [1]. Selection was done in the following steps:

- ▷ Every image class was mapped to n-dimensional vector space.

$$\mathcal{M}(class) = \mathbb{E}_{\boldsymbol{x}\sim\hat{p}_{class}} f(\boldsymbol{x};\tilde{\boldsymbol{\theta}}) \approx \frac{1}{|S_{class}|}\sum_{\boldsymbol{x}\in S_{class}} f(\boldsymbol{x};\tilde{\boldsymbol{\theta}})$$

  We used a neural network as f to map images $\boldsymbol{x}$ to vectors.
- ▷ Those vectors were clustered using k-means++.
- ▷ WordNet [2] tree was constructed for each cluster for better orientation in the database and representative classes were selected. Altogether the author selected 1150 classes.

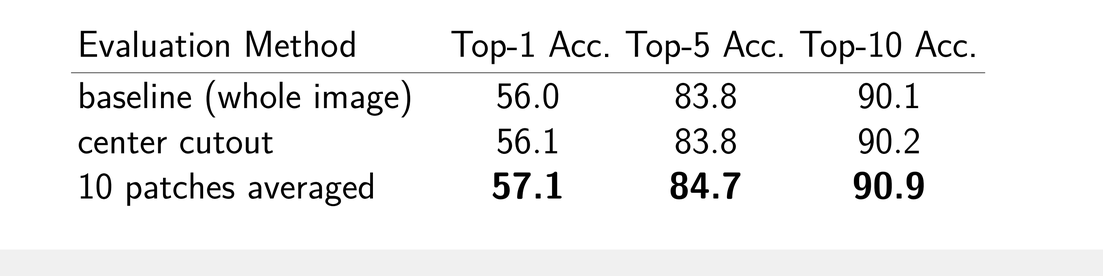

Figure 1: 2D visualization of image class vector clusters (left) and a part of WordNet tree for one cluster (right).

Retrieval Model

GoogLeNet [3] deep neural network (DCNN) was used for image annotation. We utilized transfer learning technique – the network's weights were initialized to values trained on similar task except for the last layer, which was initialized randomly. Firstly, only the last layer was trained to prevent destroying learned low-level features. Further in the training the whole network was fine-tuned.

The model was trained by 90% of the images from the selected dataset, 10% of the images were used for validation. We report accuracy on the validation set.

Evaluation Method	Top-1 Acc.	Top-5 Acc.	Top-10 Acc.
baseline (whole image)	56.0	83.8	90.1
center cutout	56.1	83.8	90.2
10 patches averaged	**57.1**	**84.7**	**90.9**

The Model at VBS 2018 Competition

The Video Browser Showdown competition [4] serves as a benchmark for interactive video retrieval tools. Our model was a part of a tool developed by SIRET group which won the competition in February 2018.

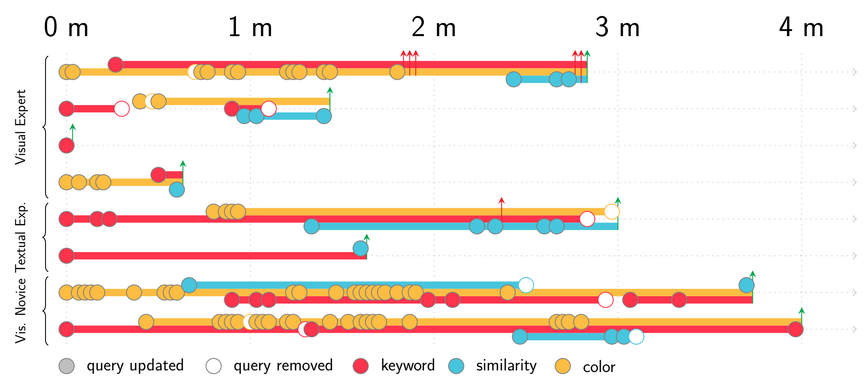

Figure 2: Use of our tool's retrieval models in some known-item search tasks at VBS 2018. The horizontal axis represents time since the task's start. We can see that our model (in red) plays an important role in success especially in the textual tasks.

Query Formulation and Ranking

The tool supports a finite set of labels a user can input. However, the labels can be connected by logical OR and logical AND to form a query

$$Q = \bigwedge_{i=1}^{k}\left(\bigvee_{\forall label_j\in N_i} label_j\right)$$

Given the query, ranking $r(\cdot)$ is assigned to each image $\boldsymbol{x}$ in a collection

$$r\left(\boldsymbol{x}; Q, \boldsymbol{\theta}\right) = \prod_{\forall N_i\in Q}\left(\sum_{\forall j\in N_i}\hat{\boldsymbol{y}}_j\cdot idf\left(j\right)\right)$$

where $\hat{\boldsymbol{y}} = f\left(\boldsymbol{x};\boldsymbol{\theta}\right)$ is the model's prediction for the image $\boldsymbol{x}$ and $idf(\cdot)$ represents an inverse document frequency (IDF) defined as

$$idf(label) = \log\left(\frac{\max_{i\in labels}\sum_{\boldsymbol{x}}\hat{\boldsymbol{y}}_i}{\sum_{\boldsymbol{x}}\hat{\boldsymbol{y}}_{label}} + 1\right)$$

Model Evaluation on KIS Task

KIS task's objective is to maximize $\mathbb{E}\left[\sum_{t=1}^{|C|} p\left(r\leq t \mid \boldsymbol{x}, Q\right)\right]$ over all possible $\boldsymbol{x}$, Q pairs given a collection C. Measuring the expectation is, however, difficult since it is dependent on the collection C and users' queries. We proposed multiple artificial users to generate $(\boldsymbol{x}, Q)$ pairs:

- ▷ **Real User.** A human judge formulates a query for a given image.
- ▷ **Network User.** Assumes coherence with the DCNN. A label for a query given an image $\boldsymbol{x}$ is randomly selected from DCNN's distribution $p(label) = \hat{\boldsymbol{y}}_{label}$.
- ▷ **Distribution User.** Given a human-generated set of image-query pairs, we inferred distribution $\mathcal{C}$ how likely user selects the top-k^{th} label as predicted by the DCNN. The top-c^{th} label from $\hat{\boldsymbol{y}}$ is added to a query where c is drawn from distribution $\mathcal{C}$.
- ▷ **Compound User.** Every query is generated by the network or by the distribution user, each user is selected with probability $1/2$.

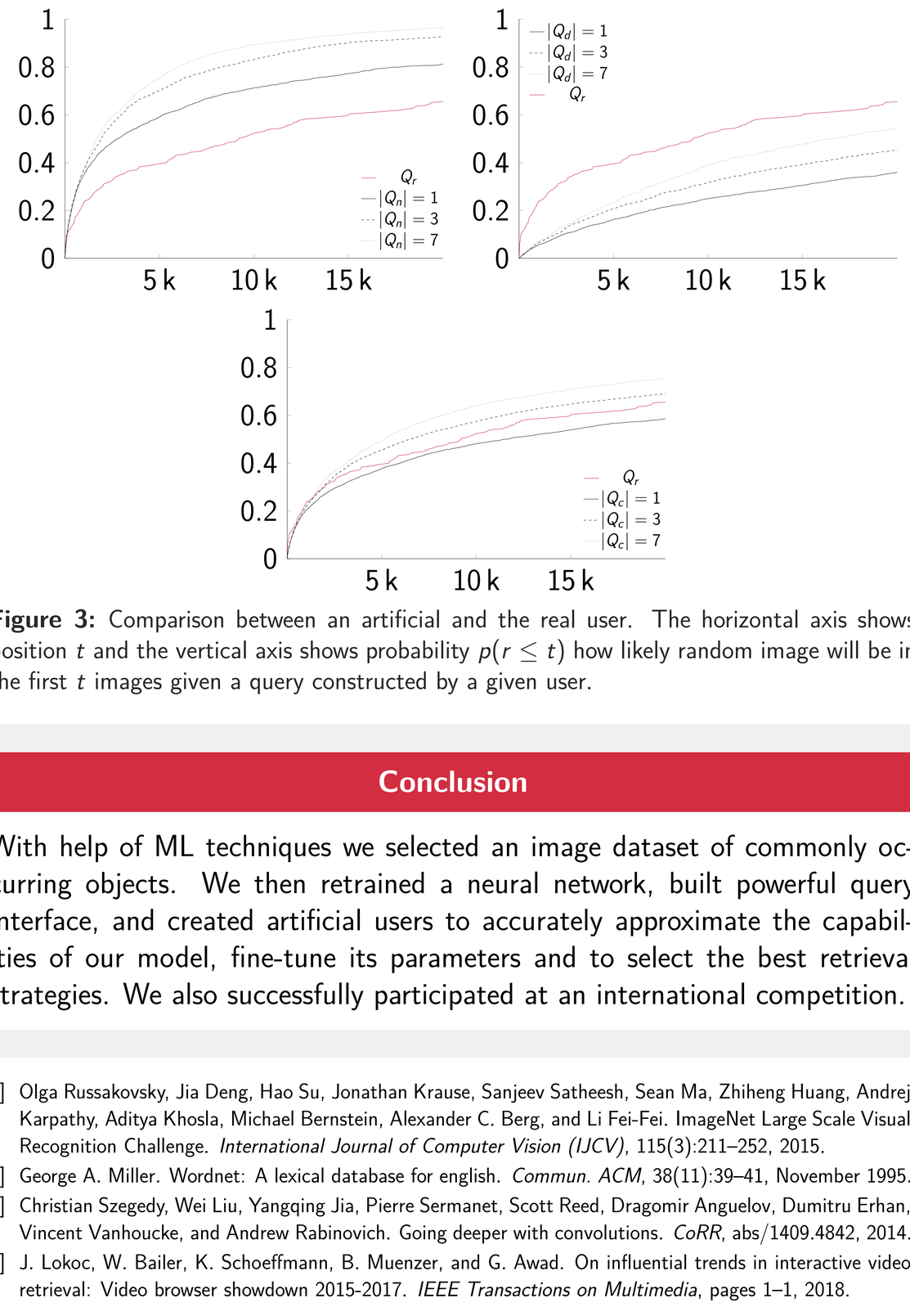

Figure 3: Comparison between an artificial and the real user. The horizontal axis shows position t and the vertical axis shows probability $p(r\leq t)$ how likely random image will be in the first t images given a query constructed by a given user.

Conclusion

With help of ML techniques we selected an image dataset of commonly occurring objects. We then retrained a neural network, built powerful query interface, and created artificial users to accurately approximate the capabilities of our model, fine-tune its parameters and to select the best retrieval strategies. We also successfully participated at an international competition.

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[2] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[4] J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad. On influential trends in interactive video retrieval: Video browser showdown 2015-2017. *IEEE Transactions on Multimedia*, pages 1–1, 2018.