Know2Look: Commonsense Knowledge for Visual Search

Sreyasi Nag Chowdhury **Niket Tandon** **Gerhard Weikum**
Max Planck Institute for Informatics
Saarbrücken, Germany
sreyasi, ntandon, weikum@mpi-inf.mpg.de

Abstract

With the rise in popularity of social media, images accompanied by contextual text form a huge section of the web. However, search and retrieval of documents are still largely dependent on solely textual cues. Although visual cues have started to gain focus, the imperfection in object/scene detection do not lead to significantly improved results. We hypothesize that the use of background commonsense knowledge on query terms can significantly aid in retrieval of documents with associated images. To this end we deploy three different modalities - text, visual cues, and commonsense knowledge pertaining to the query - as a recipe for efficient search and retrieval.

1 Introduction

Motivation: Image retrieval by querying visual contents has been on the agenda of the database, information retrieval, multimedia, and computer vision communities for decades (Liu et al., 2007; Datta et al., 2008). Search engines like Baidu, Bing or Google perform reasonably well on this task, but crucially rely on textual cues that accompany an image: tags, caption, URL string, adjacent text etc.

In recent years, deep learning has led to a boost in the quality of visual object recognition in images with fine-grained object labels (Simonyan and Zisserman, 2014; LeCun et al., 2015; Mordvintsev et al., 2015). Methods like LSDA (Hoffman et al., 2014) are trained on more than 15,000 classes of ImageNet (Deng et al., 2009) (which are mostly leaf-level synsets of WordNet (Miller, 1995)), and annotate newly seen images with class labels for bounding boxes of objects. For the image in Figure 1a, for example, object labels *traffic light, car, person, bicycle* and *bus* have been recognized making it easily retrievable for queries with these concepts. However, these labels come with uncertainty. For the image in Figure 1b, there is much higher noise in its visual object labels; so querying by visual labels would not work here.

Detected visual objects: traffic light, car, person, bicycle, bus, car, grille, radiator grille

(a) Good object detection

Detected visual objects: tv or monitor, cargo door, piano

(b) Poor object detection

Figure 1: Example cases where visual object detection may or may not aid in search and retrieval.

Opportunity and Challenge: These limitations of text-based search, on one hand, and visual-object search, on the other hand, suggest combining the cues from text and vision for more effective retrieval. Although each side of this combined feature space is incomplete and noisy, the hope is that the

“environment friendly traffic”

“downsides of mountaineering”

“street-side soulful music”

Figure 2: Sample queries containing abstract concepts and expected results of image retrieval.

combination can improve retrieval quality.

Unfortunately, images that show more sophisticated scenes, or emotions evoked on the viewer are still out of reach. Figure 2 shows three examples, along with query formulations that would likely consider these sample images as relevant results. These answers would best be retrieved by queries with abstract words (e.g. “environment friendly”) or activity words (e.g. “traffic”) rather than words that directly correspond to visual objects (e.g. “car” or “bike”). So there is a vocabulary gap, or even concept mismatch, between what users want and express in queries and the visual and textual cues that come directly with an image. This is the key problem addressed in this paper.

Approach and Contribution: To bridge the concepts and vocabulary between user queries and image features, we propose an approach that harnesses commonsense knowledge (CSK). Recent advances in automatic knowledge acquisition have produced large collections of CSK: physical (e.g. color or shape) as well as abstract (e.g. abilities) properties of everyday objects (e.g. bike, bird, sofa, etc.) (Tandon et al., 2014), subclass and part-whole relations between objects (Tandon et al., 2016), activities and their participants (Tandon et al., 2015), and more. This kind of knowledge allows us to establish relationships between our example queries and observable objects or activities in the image. For example, the following CSK triples establish relationships between *‘backpack’*, *‘tourist’* and *‘travel map’*: `(backpacks, are carried by, tourists)`, `(tourists, use, travel maps)`. This allows for retrieval of images with generic queries like *“travel with backpack”*.

This idea is worked out into a *query expansion model* where we leverage a CSK knowledge base for automatically generating additional query words. Our model unifies three kinds of features: *textual features* from the page context of an image, *visual features* obtained from recognizing fine-grained object classes in an image, and *CSK features* in the form of additional properties of the concepts referred to by query words. The weighing of the different features is crucial for query-result ranking. To this end, we have devised a method based on statistical language models (Zhai, 2008).

The paper’s contribution can be characterized as follows. We present the first model for incorporating CSK into image retrieval. We develop a full-fledged system architecture for this purpose, along with a query processor and an answer-ranking component. Our system *Know2Look*, uses commonsense *know*ledge to *look for* images relevant to a query by *looking at* the components of the images in greater detail. We further discuss experiments that compare our approach to state-of-the-art image search in various configurations. Our approach substantially improves the query result quality.

2 Related Work

Existing Commonsense Knowledge Bases: Traditionally commonsense knowledge bases were curated manually through experts (Lenat, 1995) or through crowd-sourcing (Singh et al., 2002). Modern methods of CSK acquisition are automatic, either from test corpora (Liu and Singh, 2004) or from the web (Tandon et al., 2014).

Vision and NLP: Research at the intersection of Natural Language Processing and Computer Vision is in limelight in the recent past. There have been work on automatic image annotations (Wang et al.,

2014), description generation (Vinyals et al., 2014; Ordonez et al., 2011; Mitchell et al., 2012), scene understanding (Farhadi et al., 2010), image retrieval through natural language queries (Malinowski and Fritz, 2014) etc.

Commonsense knowledge from text and vision: There have been attempts for learning CSK from real images (Chen et al., 2013) as well as from non-photo-realistic abstractions (Vedantam et al., 2015). Recent work have also leveraged CSK for visual verification of relational phrases (Sadeghi et al., 2015) and for non-visual tasks like fill-in-the-blanks by intelligent agents (Lin and Parikh, 2015). Learning commonsense from visual cues continue to be a challenge in itself. The CSK used in our work is motivated by research on CSK acquisition from the web (Tandon et al., 2014).

3 Multimodal document retrieval

Adjoining text of images may or may not explicitly annotate their visual contents. Search engines relying on only textual matches ignore information which may be solely available in the visual cues. Moreover, the intuition behind using CSK is that humans innately interpolate visual or textual information with associated latent knowledge for analysis and understanding. Hence we believe that leveraging CSK in addition to textual and visual information would take results closer to human users' preferences. In order to use such background knowledge, curating a CSK knowledge base is of primary importance. Since automatic acquisition of canonicalized CSK from the web can be costly, we conjecture that noisy subject-predicate-object (SPO) triples extracted through Open Information Extraction (Banko et al., 2007) may be used as CSK. We hypothesize that the combination of the noisy ingredients – CSK, object-classes, and textual descriptions – would create an ensemble effect providing for efficient search and retrieval. We describe the components of our architecture in the following sections.

3.1 Data, Knowledge and Features

We consider a document x from a collection X with two kinds of features:

- **Visual features xv_j:** labels of object classes recognized in the image, including their hypernyms (e.g., king cobra, cobra, snake).
- **Textual features xx_j:** words that occur in the text that accompanies the image, for example image caption.

We assume that the two kinds of features can be combined into a single feature vector $x = \langle x_1 \dots x_M \rangle$ with hyper-parameters α_v and α_x to weigh visual vs. textual features.

CSK is denoted by a set Y of triples $y_k (k = 1..j)$ with components ys_k, yp_k, yo_k (s - subject, p - predicate, o - object). Each component consists of one or more words. This yields a feature vector $y_k j (j = 1..M)$ for the triple y_k.

3.2 Language Models for Ranking

We study a variety of query-likelihood language models (LM) for ranking documents x with regard to a given query q. We assume that a query is simply a set of keywords $q_i (i = 1..L)$. In the following we formulate equations for unigram LMs, which can be simply extended to bigram LMs by using word pairs instead of single ones.

Basic LM:

$$P_{basic}[q|x] = \prod_i P[q_i|x] \quad (1)$$

where we set the weight of word q_i in x as follows:

$$P[q_i|x] = \alpha_x P[q_i|xx_j]P[xx_j|x] + \alpha_v P[q_i|xv_j]P[xv_j|x] \quad (2)$$

Here, xx_j and xv_j are unigrams in the textual or visual components of a document; α_x and α_v are hyper-parameters to weigh the textual and visual features respectively.

Smoothed LM:

$$P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1-\alpha)P[q|B] \quad (3)$$

where B is a background corpus model and $P[q|B] = \prod_i P[q_i|B]$. We use Flickr tags from the YFCC100M dataset (Thomee et al., 2015) along with their frequency of occurrences as a background corpus.

Commonsense-aware LM (a translation LM):

$$P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k]P[y_k|x]}{|k|} \right] \quad (4)$$

The summation ranges over all y_k that can bridge the query vocabulary with the image-feature vocabulary; so both of the probabilities $P[q_i|y_k]$ and $P[y_k|x]$ must be non-zero. For example, when the query asks for "electric car" and an image has features "vehicle" (visual) and "energy saving" (textual), triples such as `(car, is a type of, vehicle)` and `(electric engine, saves, energy)` would have this property. That is, we consider only commonsense triples that overlap with both the query and the image features.

The probabilities $P[q_i|y_k]$ and $P[y_k|x]$ are estimated based on the word-wise overlap between q_i and y_k and y_k and x, respectively. They also consider the confidence of the words in y_k and x.

Mixture LM (the final ranking LM):
Since a document x can capture a query term or its commonsense expansion, we formulate a mixture model for the ranking of a document with respect to a query:

$$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x] \quad (5)$$

where β_{CS} is a hyper-parameter weighing the commonsense features of the expanded query.

3.3 Feature Weights

By casting all features into word-level unigrams, we have a unified feature space with hyper-parameters (α_x, α_v, and β_{CS}). For this submission the hyper-parameters are manually chosen.

For weights of visual object class xv_j of document x, we consider the *confidence score* from LSDA (Hoffman et al., 2014). We extend these object classes with their hypernyms from WordNet which are set to the same confidence as their detected hyponyms. Although not in common parlance this kind of expansion can also be considered as CSK. We define the weight for a textual unigram xx_j as its informativeness – the inverse document frequency with respect to a background corpus (Flickr tags with frequencies).

The words in a CSK triple y_k have non-uniform weights proportional to their similarity with the query words, their *idf* with respect to a background corpus, and the salience of their position – boosting the weight of words in s and o components of y. The function computing similarity between two unigrams favors exact matches to partial matches.

3.4 Example

Query string: *travel with backpack*
Commonsense triples to expand query:
$t1$: `(tourists, use, travel maps)`
$t2$: `(tourists, carry, backpacks)`
$t3$: `(backpack, is a type of, bag)`

Say we have a document x with features:
Textual - "A tourist reading a map by the road."
Visual - person, bag, bottle, bus

The query will now successfully retrieve the above document, whereas it would have been missed by text-only systems.

4 Datasets

For the purpose of demonstration we choose a topical domain – *Tourism*. Our CSK knowledge base and image dataset obey this constraint.

CSK acquisition through OpenIE: We consider a slice of Wikipedia pertaining to the domain *tourism* as the text corpus to extract CSK from. Nouns from the Wikipedia article titled 'Tourism'(seed document) constitute our basic language model. We collect articles by traversing the Wiki Category hierarchy tree while pruning out those with substantial topic drift. The Jaccard Distance (Equation 6) of a document from the seed document is used as a metric for pruning.

$$JaccardDistance = 1 - WeightedJaccardSimilarity \quad (6)$$

where,

$$WeightedJaccardSimilarity = \frac{\Sigma_n min[f(d_i, w_n), f(D, w_n)]}{\Sigma_n max[f(d_i, w_n), f(D, w_n)]} \quad (7)$$

In Equation 7, acquired Wikipedia articles d_i are compared to the seed document D; $f(d', w)$ is the frequency of occurrence of word w in document d'. For simplicity only articles with Jaccard distance of 1 from the seed document are pruned out. The corpus of domain-specific pages thus collected constitute ~5000 Wikipedia articles.

Table 1: Query Benchmark for evaluation

aircraft	international	diesel	transport
airport	vehicle	dog	park
backpack	travel	fish	market
ball	park	housing	town
bench	high	lamp	home
bicycle	road	old	clock
bicycle	trip	road	signal
bird	park	table	home
boat	tour	tourist	bus
bridge	road	van	road

The OpenIE tool ReVerb (Fader et al., 2011) run against our corpus produces around 1 million noisy SPO triples. After filtering with our basic language model we have ~22,000 moderately clean assertions.

Image Dataset: For the purpose of experiments we construct our own image dataset. ~50,000 images with descriptions are collected from the following datasets: Flickr30k (Young et al., 2014), Pascal Sentences (Rashtchian et al., 2010), SBU Captioned Photo Dataset (Ordonez et al., 2011), and MSCOCO (Lin et al., 2014). The images are collected by comparing their textual descriptions with our basic language model for *Tourism*. An existing object detection algorithm – LSDA (Hoffman et al., 2014) – is used for object detection in the images. The detected object classes are based on the 7000 leaf nodes of ImageNet (Deng et al., 2009). We also expand these classes by adding their super-classes or hypernyms with the same confidence score.

Query Benchmark: We construct a benchmark of 20 queries from co-occurring Flickr tags from the YFCC100M dataset (Thomee et al., 2015). This benchmark is shown in Table 1. Each query consists of two keywords that have appeared together with high frequency as user tags in Flickr images.

5 Experiments

Baseline Google search results on our image dataset form the baseline for the evaluation of *Know2Look*. We consider the results in two settings – search only on original image caption (Vanilla Google), and on image captions along with detected object classes (Extended Google). The later is done to aid Google in its search by providing additional visual cues. We exploit the domain restriction facility of Google search (*query string site:domain name*) to get Google search results explicitly on our dataset.

Table 2: Comparison of *Know2Look* with baselines

	Average Precision@10
Vanilla Google	0.47
Extended Google	0.64
Know2Look	0.85

Know2Look In addition to the setup for Extended Google, *Know2Look* also performs query expansion with CSK. In most cases we win over the baseline since CSK captures additional concepts related to query terms enhancing latent information that may be present in the images. We consider the top 10 retrieval results of the two baselines and *Know2Look* for the 20 queries in our query benchmark[1]. We compare the three systems by Precision@10. Table 2 shows the values of Precision@10 averaged over 20 queries for each of the three systems – *Know2Look* performs better than the baselines.

6 Conclusion

In this paper we propose the incorporation of commonsense knowledge for image retrieval. Our architecture, *Know2Look*, expands queries by related commonsense knowledge and retrieves images based on their visual and textual contents. By utilizing the visual and commonsense modalities we make search results more appealing to the humans than traditional text-only approaches. We support our claim by comparing *Know2Look* to Google search on our image data set. The proposed concept can be easily extrapolated to document retrieval. Moreover, in addition to using noisy OpenIE triples as commonsense knowledge, we aim to leverage existing commonsense knowledge bases for future evaluations of *Know2Look*.

Acknowledgment: We would like to thank Anna Rohrbach for her assistance with visual object detection of our image data set using LSDA. We also thank Ali Shah for his help with visualization of the evaluation results.

[1] http://mpi-inf.mpg.de/~sreyasi/queries/evaluation.html

References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction for the web. In *IJCAI*, volume 7, pages 2670–2676.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer.

Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. 2014. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Xiao Lin and Devi Parikh. 2015. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2984–2993.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. 2007. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690.

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

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. 2015. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1456–1464. IEEE.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the move to meaningful internet systems 2002: Coopis, doa, and odbase*, pages 1223–1237. Springer.

Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 523–532. ACM.

Niket Tandon, Gerard de Melo, Abir De, and Gerhard Weikum. 2015. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proc. CIKM*.

Niket Tandon, Charles Hariman, Jacopo Urbani, Anna Rohrbach, Marcus Rohrbach, and Gerhard Weikum. 2016. Commonsense in parts: Mining part-whole relations from the web and image tags. *AAAI*.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*.

Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2542–2550.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

Josiah K Wang, Fei Yan, Ahmet Aker, and Robert Gaizauskas. 2014. A poodle or a dog? evaluating automatic image annotation using human descriptions at different levels of granularity. *V&L Net 2014*, page 38.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

ChengXiang Zhai. 2008. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141.