Towards Augmented Database Schemes by Discovery of Latent Visual Attributes

Visionary

Tomáš Grošup, Ladislav Peška, Tomáš Skopal
SIRET Research Group, Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic
{grosup,peska,skopal}@ksi.mff.cuni.cz

ABSTRACT

When searching for complex data entities, such as products in an e-shop, relational attributes are used as filters within structured queries. However, in many domains the visual appearance of an item is important for a user, while coverage of visual appearance by relational attributes is left to database designer at design time and is by nature an incomplete and imperfect representation of the entity. Recent advances in computer vision, dominated by deep convolutional neural networks (DCNNs), are a promising tool to cover the gaps. It has been shown that activations of neurons of DCNNs correspond to understandable visual-semantic features of an input image. We envision that activations of neurons are of great use for search queries in domains with strong visual information, even when obtained from DCNNs models pre-trained on general imagery. Locally scoped visual features obtained using them can be combined to form search masks which would correlate to what humans understand as an attribute, when applied on the entire dataset. Ultimately, combination of visual features can be identified automatically and formed into immediate suggestion of a new relational attribute, leaving one last task for humans to turn this into augmentation of the database schema – putting a label on it.

1 INTRODUCTION

Many approaches to information retrieval tasks, e.g., querying or exploration, assume structured data entities. These are modelled using composition of simple attributes and relations to other structured entities and as such, they represent a simplified model of the real world entity. Object's attributes either comply to a fixed schema in relational database model, or there is an implicit schema defined by existence of attributes in non-relational models such as document databases, key-value stores or linked data. Even though some data representations (e.g., JSON, XML) might be referred as schema-less and compliance is not always enforced, applications are working against an expected structure of the data and their attributes. In either case, these attributes are defined in a supervised manner by domain experts.

As an example application, let us consider an e-shop with a search interface. Within its search interface, users can define structured queries, which operate on a pre-defined set of attributes, utilizing range filters, exact matching or similar filtering mechanisms. An example query might be "*category = 'shoes' AND price between 100 and 200 AND color = 'black' AND description matches 'summer'*". Usability of the querying mechanism is limited both by the expressiveness of the database schema as well as capability of the user. In many domains (e.g., art, cars, dating, decorations, fashion or furniture), schema expressiveness is rather inferior due to the low information gain of available structured data. Therefore, the final filtering step of a retrieval task has to be executed by a human via examination of non-structured data (images, audio, video, etc). Scientific communities have been trying to reveal the internal structures of multimedia content in the last decades with gradually improving results. However, disclosing the multimedia object's representation in a way comprehensible for end-users remains extremely challenging. Partial solutions of this challenge are query-by-example search or multimedia exploration paradigms, accompanying the classical structured search as a refinement step.

Recent advances in neural network architectures, such as deep convolutional neural network AlexNet [8] represent a promising direction in revealing the internal structure of multimedia. Due to their layered nature, learned concepts are more high-level, and can bridge the semantic gap in computer vision. Although the task of many pioneer architectures was to classify data into a fixed set of pre-defined categories, it has been shown that the trained models can generalize to unknown domains, if the activation of neurons is used as a high-dimensional descriptor of objects [3]. Existing network models can also be fine-tuned to operate on a finer subset of data, increasing the precision for a particular dataset [4]. Finally, a layered networks' architecture allows to generate descriptors on different levels of a semantic scale [13, 19], ranging from low-level visual features (e.g., edges) up to high-level concepts, such as a cat. It is also possible to focus on individual regions of an image, denoted as patches [5, 18].

In our vision, we propose to extract information encoded in these descriptors via identification of commonly occurring combination of features. In order to analyze the behaviour of descriptors in the domain of product imagery, we implemented a multi-example similarity search mechanism [15]. This was also used to better understand applicability of different layers of the network by evaluating precision of their descriptors against manually selected proposals of new attributes acting as a test dataset. In this first step, the similarity was evaluated only globally, by extracting and comparing descriptors representing the whole object. As a second step, evaluation of individual image parts, patches, was done [14]. Human users were tasked to manually highlight interesting part of the image when selecting a set of images as an input for the search. Selected parts of an image were powering the search mechanism, which operated on an aggregation of global representations of images as well as on individual descriptors of each image patch, matching patches in the input query and in the overall database. This was partly done as an analytical step to better understand the behaviour and patterns within the data and how they correlate to what humans understand as an attribute.

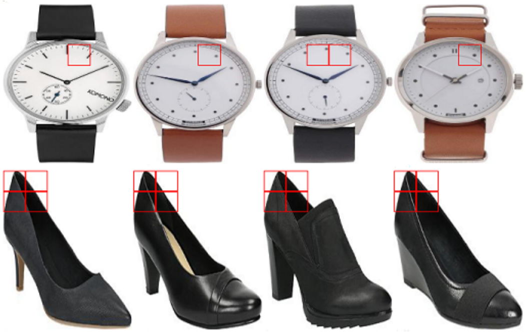

Figure 1: Examples of two proposed attributes. First row are "hand watches without labels". Second row are "lady shoes with sharp corner in the ankle area".

In the final stage of our vision, patterns of descriptors occurring frequently in the dataset can be automatically suggested as candidates for new attributes, immediately augmenting the original database schema (see Figure 1). Once the schema is augmented, all of the existing applications using it as a building block, e.g., recommender systems, data analytics, search functions and many others, can utilize this augmentation and therefore cummulatively increase the added value of the newly discovered "virtual" attribute. This implies benefits from an overall data-engineering perspective, and not just an application-specific improvement of a search function.

2 RELATED WORK

To the best of our knowledge, there have been no attempts to augment database schema using multimedia features. However, there are many similar disciplines that were used as inspiration and many technologies that could be used as building blocks to help implementing the vision. We want to highlight that rather than solving a single application-specific task (e.g., using images for product recommendation) we define the problem from database-engineering perspective, so that each service utilizing a database schema could benefit from its refinement.

Deep learning approaches have been evolving in the last years. Recent advances include inception modules (Googlenet), residual networks (ResNet, ResNeXt, DenseNet), and meta-learning approaches for searching an optimal architecture automatically (NAS), with better architectures being proposed every year. The target of our vision is to work with any network pre-trained on general imagery, making the choice of network architecture orthogonal to the main problem.

One of anticipated outcomes of our vision is an improvement of product recommendation techniques. Several papers focusing on fashion recommendation were published recently [6, 7, 9, 12, 17]. In fashion domain, visual-similarity based recommending approaches managed to considerably improve conversions of visitors to buyers [12]. In [6, 7, 12], authors focused on similarity and identity matching of "wild" images (i.e., pictures taken on streets or non-canonical representations of objects), while our scenario allows us to restrict on canonical representations of objects shot from similar angles with uniform background, etc.

Somewhat similar to our work is the approach by Yu et al. [17] on utilization of aesthetic-based features of objects together with common CNN features. However, using sole aesthetic features did not improve over CNN features and therefore we kept focusing on local visual similarity of objects instead of their aesthetics. The main difference between our vision and mentioned related work is the aim on materializing relevant similarity pattern instead of just utilizing them as a part of similarity calculations.

The topic of schema augmentation is not new, however, it has not been targeted via latent visual information so far. Selke et al. [11] focused on crown-enabled databases, which could enrich database schema at the query-time by utilization of manual labor to power search execution. Yakout et al. [16] automatically augmented entities via scraping of online available HTML tables and matching their information.

3 ROAD MAP

In this section, we overview the main points of our vision, describe the context, where it brings the most benefit, and define individual milestones and already completed goals, which altogether form a road map towards turning the vision into reality.

3.1 Vision

The body of our vision is to provide global as well as a user-based schema extensions by virtual attributes not present in the original database schema. Virtual attributes are to be derived from a visual information in domains, where it can provide a significant information gain, e.g., fashion, art, decorations, furniture, etc. We do not meant to replace classical search or recommendation methods, but rather boost them by providing additional relevant features. In addition to the applications in automated content processing methods, virtual attributes (if properly labeled) can be revealed to the end-users or utilized in data analytics.

Implicit feedback collected in an application-specific scope, such as browsing history in an e-shop, can be used as crowd-based evidence for the suggested attributes. If an attribute correlates with contents of search history, it is likely to have a meaning to humans. The evidence could naturally grow from single random visitors, over long-term users, up to large groups of similar users and thus increasing the level of confidence for the suggestion.

A particular objective aims at straightforward applicability for general usage; not relying on extensive training phase and pre-existing definitions of attributes/classes. The main challenge relies in an operation on previously unseen and undefined attributes, and an evolution of the model at runtime. Generally, training images and classes are essential for today's state-of-the-art computer vision models such as neural networks. However, the real world has unlimited number of classes assignable to objects, and it is impossible to prepare such data upfront. At the same time, requiring a special design and training phase would implicitly mean a need for a trained machine learning specialist, which limits the benefits to a small portion of data owners. We believe that it is possible to finalize the envisioned method in a set of unsupervised-pipeline steps, leading to a deployment-enabled black-box component which can be integrated into a broad range of today's applications.

3.2 Context

The envisioned approach of automatically detecting attributes is not guaranteed to be beneficial within all domains, not even when restricting the idea to products with visual information. The main preconditions are that visual information is an important factor influencing users' decisions and that information gain of existing relational attributes is low compared to information retrieved from multimedia data. This condition disqualifies product domains such as computer components, where the visual information is available, but existing relational attributes cover most of the users' information needs.

The applicability of our approach rises in scenarios, where objects may contain multiple important visual attributes and therefore it is not possible to simply label objects w.r.t. some class hierarchy. It is the dynamic point of view making it a challenge, since different users looking at the exact same image might have different understanding what is the attribute of interest for them. For example, several virtual attributes, e.g., leather-build, high heel, floral texture and black zipper may be detected for a single shoe, however for various users, only some of the attributes may be relevant in their decisions.

3.3 Milestones

In the following we summarize the milestones of the road map, also sketched in Figure 2.

Descriptor extraction (a). As a first step of our vision, proper image descriptors need to be selected. However, due to the broad range and complexity of today's techniques, it is almost impossible to evaluate all possible variants and combinations. Even with simple DCNN like AlexNet, there are several layers to choose from and different aggregation methods to turn convolutional layers into a single vector. We have evaluated performance of different layers in [15] with the conclusion that there is no single dominating layer for all search tasks. Therefore, this aspect still deserves more research attention. Nonetheless, as all further tasks are independent on the choice of image descriptors, it is plausible to start with some simple selections, e.g., based on AlexNet and research this area in parallel with further tasks.

Image patches (b) are regions of an image that enable to focus on a particular part of it and, as a consequence, evaluate its similarity locally. Within standardized product datasets, it can be sufficient to cut images using regular grids, since all images have same orientation, centering and background. If we move towards "images in the wild" scenario, some trainable image segmentation approach might be necessary. In either case, the final image representation is a list of patches' descriptors. Another challenge in this task is to define methods to aggregate similarity w.r.t. different patches. We already reported results of some basic aggregation methods in [14].

Similarity sets (c), defined by image patches that are similar to each other in a given feature-space, could be the first step towards a virtual attribute. There are multiple ways to model similarity between image patches, e.g., by evaluating only the distance between their respective descriptors or by weighting it together with global similarity of the entire entity.

Noise removal (d) in the area of similarity sets is a challenging problem given the unsupervised nature of the desired pipeline, i.e., the only available information are images, their descriptors, and similarity values. The distribution of distance values can be a useful heuristic to filter out sets which are unlikely to form a good attribute, such as near-duplicates (distance close to zero), trivial patches (small distance to single-color patches) or sets being too large (attributes would no longer be discriminatory). Another possibility is to eliminate noise by cross-checking data coverage by existing relational attributes, e.g., to remove suggestion for an attribute that is already known in the data and correlates highly with the new suggestion.

Frequent patterns (e) within the filtered similarity sets should be prioritized in order to provide a ranked list of suggested attributes to the domain administrators. Existing approaches to the frequent patterns mining, such as the Market Basket Analysis [1], could be utilized. The algorithm should be optimized to reduce the number of suggested patterns/rules, while preserving a good coverage of entities in the database and a reasonable confidence level that can be used for sorting purposes.

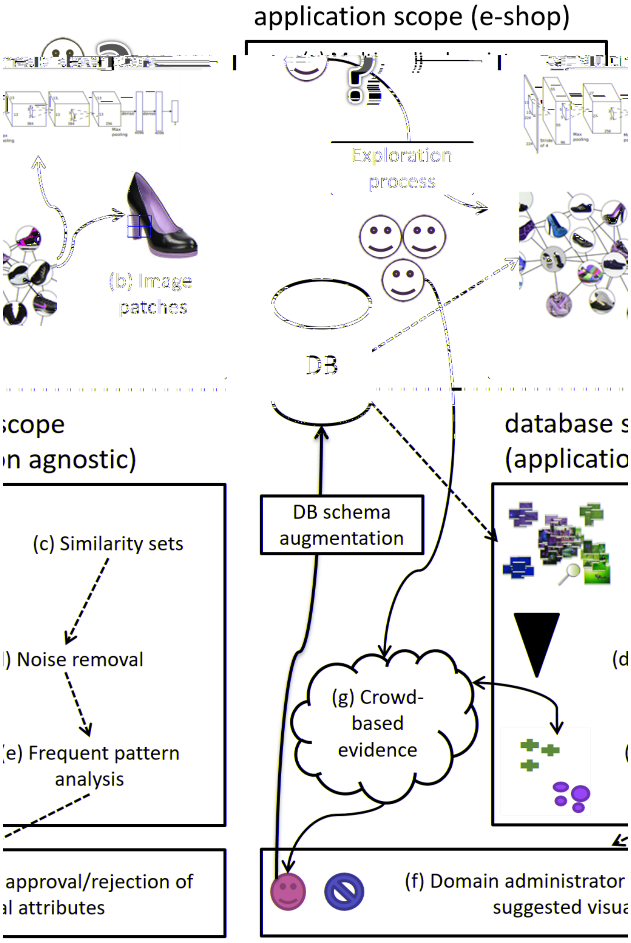

Figure 2: Road map/architecture of the vision.

Domain administrator (f) could be presented with n-tuples of items that match a suggested attribute. Furthermore, if the feedback from users is available (e.g., their browsing history or shopping baskets), frequently co-occurring items possessing a suggested attribute may be prioritized. The interface could also highlight the features on which the similarity of the set is based, e.g., by highlighting relevant image patches. If the suggestion gets approved and labeled, the similarity set could be turned into a search mask, evaluated w.r.t. the full database of descriptors to turn it into a new attribute. Since the database of descriptors contains also individual patches, certain attributes could be detected multiple times within a single object (a 1:N relation).

Collaborative techniques (g) based on implicit user feedback could not only refine suggestions to the domain administrators, but could also establish sets of latently approved virtual attributes specific for a cluster of users, or a single long-term user. In such a way, we may postpone the work of domain administrators and label the attributes only upon a request from a component that disclose attributes to the users (e.g., data analytics or attribute search). Challenges of this task are balancing the granularity of collected feedback (e.g., a single user, top-k similarity, user clustering), maintaining different augmented schemes for multiple user scopes and proposing methods capable to aggregate more granular augmented schemes (e.g., on a level of single users) and bring them to the higher level, eventually reaching the main database schema.

3.4 Challenges

In this section, we summarize the major challenges of the proposed vision and offer ideas on how to address them. The challenges are evaluation, noise reduction, scalability, personalization, continuous schema evolution and applicability on other forms of content.

By far the most significant challenge for database augmentation is the lack of an established evaluation framework and standardized datasets. The nature of human-system interaction will require time-consuming user studies in order to collect sufficient training data and feedback. The evaluation metrics could measure the overall user satisfaction, impact on speed in information retrieval tasks, as well as indirect impact obtained due to improved recommendation and search, e.g., the conversions ratio. Also, the practically unbound volume of possible attribute suggestions and the lack of training data for ranking the candidates makes human-in-the-loop a critical part of the evaluation.

The second challenge is the reduction of noise. The amount of attributes within a dataset is subject to a combinatorical explosion of common patterns across the dataset. The system can verify their meaningfulness using automated techniques operating on similarity data and visual descriptors, but it might be very difficult to systematically estimate if a pattern is indeed a new attribute, or if it was a random set of data points.

In order to deal with a large volume of possible patterns, the system must be efficient and scalable in all parts of data processing pipeline. In the existing work, the MapReduce paradigm was used in the implementation, allowing the computation to run on a large cluster of machines. The computationally intensive step of similarity self-join identifying common patterns across the dataset, was executed using the Hadoop MapReduce algorithms of Čech et al. [2].

The fourth challenge is personalization. Whenever the system starts inferring new attributes based on the implicit feedback input, the correct scope must be identified. The trivial cases are single-user scoped and globally scoped attributes. However, collaborative approaches may consider various, possibly overlapping clusters of similar users. This would require more extensive verification of an attribute across different user clusters, in order to estimate its benefit for the respective user groups.

The fifth challenge is the continuous schema evolution, fostered by a stream of new attributes, and its propagation into classical database tasks, e.g., similarity search. Therefore, the inclusion of new attributes must be done in a transparent way, so the successive techniques can automatically incorporate it without explicit intervention. Transitively, this has also impact on all other components operating on a set of relational attributes, such as data exploration or recommendation engine.

The last challenge is to generalize the proposed model into other forms of multimedia, such as sound or video. This might be appealing in domains, where short video material represents an inherent part of information retrieval tasks. One example might be a database of movies with their trailers. Recently, there have been significant improvements in video browsing techniques, such as the work of Lokoč et al [10]. The proposed schema augmentation approach could be beneficial in such scenarios as well.

4 CONCLUSION

This vision paper outlines a novel research problem, which aims at augmenting database schemes by attributes extracted from visual information. Initial attempts have outlined a possible direction for future research and identified several sub-problems and challenges to solve.

Acknowledgments

This research has been supported by Czech Science Foundation (GAČR) project Nr. 17-22224S.

REFERENCES

[1] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. 1997. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD '97)*. ACM, New York, NY, USA, 255–264.

[2] Přemysl Čech, Jakub Maroušek, Jakub Lokoč, Yasin N. Silva, and Jeremy Starks. 2017. Comparing MapReduce-Based k-NN Similarity Joins on Hadoop for High-Dimensional Data. In *Advanced Data Mining and Applications*. Springer, 63–75.

[3] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *CoRR* abs/1310.1531 (2013). arXiv:1310.1531

[4] Z. Ge, C. McCool, C. Sanderson, and P. Corke. 2015. Subset feature learning for fine-grained category classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 46–52.

[5] Xufeng Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. 2015. MatchNet: Unifying feature and metric learning for patch-based matching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3279–3286.

[6] J. H. Hsiao and L. J. Li. 2014. On visual similarity based interactive product recommendation for online shopping. In *2014 IEEE International Conference on Image Processing (ICIP)*. 3038–3041. https://doi.org/10.1109/ICIP.2014.7025614

[7] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. 2015. Where to Buy It: Matching Street Clothing Photos in Online Shops. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3343–3351.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 1097–1105.

[9] Nick Landia. 2017. Building Recommender Systems for Fashion: Industry Talk Abstract. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, 343–343. https://doi.org/10.1145/3109859.3109929

[10] J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad. 2018. On influential trends in interactive video retrieval: Video Browser Showdown 2015-2017. *IEEE Transactions on Multimedia* (2018).

[11] Joachim Selke, Christoph Lofi, and Wolf-Tilo Balke. 2012. Pushing the Boundaries of Crowd-enabled Databases with Query-driven Schema Expansion. *Proc. VLDB Endow.* 5, 6 (Feb. 2012), 538–549.

[12] Devashish Shankar, Sujay Narumanchi, H A Ananya, Pramod Kompalli, and Krishnendu Chaudhury. 2017. Deep Learning based Large Scale Visual Recommendation and Search for E-Commerce. (2017). arXiv:1703.02344

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR* abs/1312.6034 (2013).

[14] Tomás Skopal, Ladislav Peska, and Tomás Grosup. 2018. Interactive Product Search Based on Global and Local Visual-Semantic Features. In *Similarity Search and Applications - 11th Int. Conference, SISAP 2018, Lima, Peru*. 87–95.

[15] Tomáš Skopal, Ladislav Peška, Gregor Kovalčík, Tomáš Grosup, and Jakub Lokoč. 2017. Product Exploration Based on Latent Visual Attributes. In *Proc. of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, 2531–2534. https://doi.org/10.1145/3132847.3133175

[16] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *Proc. of the 2012 ACM SIGMOD Int. Conference on Management of Data (SIGMOD '12)*. ACM, New York, NY, USA, 97–108.

[17] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based Clothing Recommendation. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. Int. World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 649–658.

[18] Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to Compare Image Patches via Convolutional Neural Networks. *CoRR* abs/1504.03641 (2015).

[19] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 818–833.