Mladen Fernežir, Enes Deumić, Ivan Borko, Vjekoslav Giacometti, Dominik Šafarić, Marko Velić, Vedran Vekić, Davor Aničić

Computer Vision R&D for Classifieds in Styria Media Group

Styria Media Services ltd. Data Science Department, Oreškovićeva 6H/1, 10000 Zagreb, Croatia

Abstract

In this paper, we present two computer vision projects that were deployed as services for the Styria Media Group's classifieds: hierarchical fine-grained image categorization and image similarity search. For image categorization, we generalize the previous accuracy vs. specificity approach to automatically offer sets having the best combined accuracy and specificity, instead of returning single element suggestions. We also modify the original specificity measure to be more appropriate for the classifieds use case: minimizing the number of required clicks to reach the desired leaf category. Further, we describe our approach of utilizing a deep learning classification model for another task: creating binary descriptors in an end-to-end manner to be used for image similarity retrieval. To accomplish this task, we combine various features from different parts of the network, use multimodal learning which combines images and text from classified's ads, and finally, we employ triplet metric learning for color encoding.

1. Introduction

Styria, founded in 1869, is one of the leading media groups in Austria, Croatia, and Slovenia. As a part of the Styria Media Group, in early 2015, a team was formed to develop data science solutions for the entire group, combining natural language processing and computer vision research. Computer vision research and development for the Classifieds Project started with a clear goal to improve user experience on both the buyers' and the sellers' side of the online sales process for the Styria Group's classifieds (2nd hand marketplaces). The goal was to encourage users to do more ad placements and to have more productive searches. This would directly increase the value of the classified for its users.

For the buyers' side, the result of the project is a service called Fashion Cam, built for the Austrian Willhaben classified. The service enables buyers to find visually similar objects more easily. At first, the service was developed only for fashion but now also for furniture and antiques, with other categories soon to follow.

For the sellers' side, the end result is automatic category suggestion based on one or more images, developed for the Njuskalo classified in Croatia. The service makes the ad posting process easier and faster for the sellers.

Both products were possible due to recent advances in deep learning [1], [2], specifically in Convolutional Neural Networks (CNNs) [3]. The progress in the field was facilitated by the availability of large amounts of labeled data, modern GPU advancements, and also by hosting large-scale visual recognition competitions in the academic community based on the ImageNet dataset [4].

2. Hierarchical fine-grained image categorization

For the classifieds use case, it is common for the categories to be organized in a hierarchical manner into a specific category tree. Typically, there are multiple problems to handle: semantically similar categories in different parts of the tree, highly uneven category distributions, label quality concerns, and also, issues related to the fine-grained nature of objects to be recognized. For such fine-grained use cases, there is a problem of large intra-class variance and at the same time, small inter-class variance between some categories in the classified's categorization tree. The fine-grained problem is an active area of research tackled on diverse datasets, e.g. Oxford Flowers [6], Oxford-IIIT Pet [7], Stanford Dogs [8], CUB200-2011[9] and Cars196 [10].

At first, the problem was approached as a standard leaf classification task. The CNN network was trained to predict confidences for each of the leaf categories, using the actual leaf categories that users had chosen when placing the ads as ground truth labels for each input image. For cases where there were multiple images for the same ad, confidence predictions were averaged to obtain more accurate results.

To return the final category suggestion to our client, a separate model was trained to suggest the best subset of up to 3 nodes in the classified's categorization tree.

2.1. Architectures

The choice of the actual CNN architecture is determined by two factors: actual classification performance, and also by the required computational performance to be able to handle real-time classification requests. Currently, the models in production use elements of the GoogLeNet [11], Darknet[12] and DenseNet[13] architectures.

2.2. Revisiting the accuracy-specificity trade-off

When dealing with a hierarchical category structure, there is a possibility of returning one or more inner nodes in the categorization tree as the final category suggestion, instead of just the most confident leaf. This enables gains in accuracy, at the cost of some specificity.

Our initial solution was adopted from the Hedging Your Bets paper [14]. The paper defines a measure of specificity for each of the tree nodes, which enables joined confidence and specificity node scoring. The final category suggestion is the node having the maximum score.

Still, since there are cases of semantically similar categories in distant parts of the categorization tree, in many cases it would be best to offer all similar suggestions. The main limitation of the original HYB algorithm is that it could only offer one node suggestion. This would result in missing some of the legitimate suggestions, or moving back all the way to the common ancestor too close to the root of the categorization tree.

To solve these issues, we redesigned the original algorithm to generalize scoring to sets of nodes. This required a redefined specificity measure, which was also more appropriate for the final use case: minimizing the number of clicks that the user would have to take from our suggestions to the desired leaf category.

2.3. Categorization examples

Figures 3. and 4. showcase our category suggestions. Note that in the first case (Figure 3.) "hand tools" appear at two different places in the categorization tree. The second case illustrates a typical situation when it makes sense to offer both men's and women's categories (Figure 4).

3. Custom and fast visual search for real world images

For the Fashion Cam service and image similarity search in general, the biggest problem is the definition of similarity itself. There is always a semantic component, corresponding to the classified's leaf category the ad was placed in. Other aspects are more visual: material, shape, texture, and color. In some cases, there is also the brand component which has its own important semantic and visual contributions.

The end product had to take into account both semantic and visual aspects when returning the most similar image for a given image query. At the same time, it also had to be fast to offer real-time service to our clients. Another limitation was in the available data itself which only had ad category annotations, without additional attribute tags.

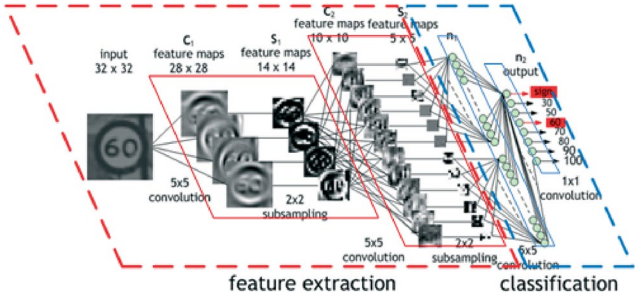

Fig. 1. Convolutional neural networks enable hierarchical learning of features: from more basic like edges and blobs to more abstract ones, enabling final object categorization. Image by Maurice Peemen [5].

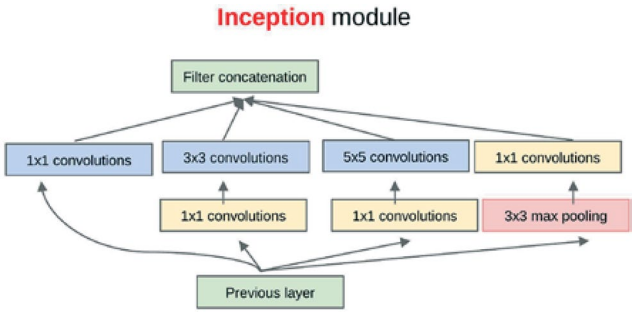

Fig. 2. Inception module, the basic component of the GoogLeNet architecture. The input layer is examined by convolutions of different kernel sizes (1 x 1, 3 x 3 and 5 x 5).

Fig. 3. Suggested tree nodes: 1. Machine and tools / Construction machinery and tools / Hand tools and tools; 2. Machine and tools / Hand tools

Fig. 4. Suggested tree nodes: 1. Fashion / Apparel / Watches / Men's watch; 2. Fashion / Apparel / Watches / Smartwatch; 3. Fashion / Apparel / Watches / Women's watch

Fig. 5. Search results when using image descriptors more focused on semantics (top row) and when using descriptors with more emphasis on visual features (bottom row).

The approach we used to solve the image similarity search problem falls into the general category of representation learning [15], and more specifically, into the category of searching the appropriate hashing representation for each image with a data-dependent approach. An overview of the most recent data-dependent approaches to hashing is provided in [16]. We use a deep learning based data-dependent approach for two reasons: utilizing all specifics in the data to obtain better descriptors, and to have a fast end-to-end solution ready for real-time service for our clients.

3.1 Descriptor extraction and binary encoding

The first model followed the idea presented in [17] to train a binary descriptor designed to capture category level semantics. They added an extra sigmoid fully connected layer in-between the final feature layer and the logits layer used for classification, with the idea to train that layer so that it captures high-level semantics. Two additional training loss components were used: one to make sigmoid activations close to 0 and 1, and another to make the activations as diverse as possible.

This solution was a good starting point to capture category semantics. However, it was soon discovered that we would have to do better to capture more visual aspects, especially for the fashion use case. Also, unlike [17] that used the binarized sigmoid layer for a first, coarse-level search and still reverted to a large float descriptor for fine semantic comparison, we desired a fully binarized solution to meet our run-time requirements.

To accomplish these goals of encoding both visual and semantic aspects, and to have a fully binarized descriptor, we investigated other layers in the deep neural network besides the top one meant for semantics. We took advantage of the nature of deep learning with convolutional neural networks that was mentioned in the introduction: the network learns the needed concepts hierarchically, from simpler to more abstract. The more visual aspects were present in the lower parts of the network. The final binary descriptor was formed from many different parts of the network, with many tweaks to get the satisfactory balance of semantics and visuality. Figure 5. illustrates two different combinations: a more semantically based one and a more visual one.

For faster run-time, we used a simple fully connected autoencoder to encode the final binary descriptor into a smaller one of size 64. The small one is meant for the first coarse-level search and the full one for the final ranking. All comparisons are fast on modern CPU architectures since the Hamming distance (Figure 6.) between binary descriptors can be calculated by simple XOR and bit count operations.

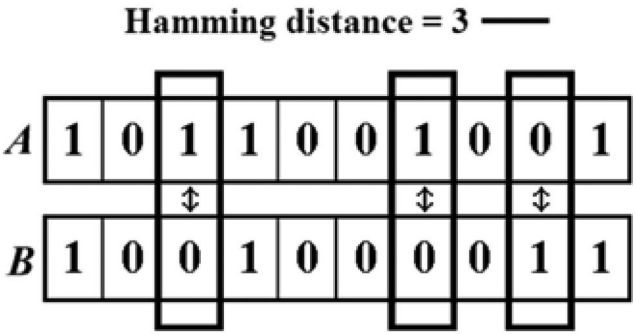

Fig. 6. The Hamming distance calculates the number of differing bits between two binary descriptors.

3.2 Color encoding

Color is a visual aspect that was especially important for our users. To enhance their experience, we trained a separate color encoding model and injected the color encoding layers into the main network for an end-to-end run time solution. We used triplet metric learning [18] to map perceptually similar colors in CIELAB color space to binary descriptors having similar Hamming distances.

3.3 Detecting brands

For some categories, it was especially important to be able to retrieve objects which correspond to the same brand as the query image. To accomplish this, we used a multimodal deep learning approach [19]. We used tex-

Fig. 7. Search results where brand retrieval was especially important.

tual information from the ads to detect most informative words with respect to the category in which the ad was placed. In many cases, these were brands along with some other typical words that represent types of materials. After that, the network was re-trained with this information to serve as an additional goal for learning. Results turned out to be quite good, especially for categories like sneakers or women's purses. Figure 7. illustrates similarity search results for men's Nike sneakers

4. Project results

Our response times are around 100 ms for categorization and search-by-image services, and just 50 ms when using an image that is already present on the site as a search query.

The time spent by the user in the ad insertion process, from the click on "post a new ad" until inputting text, was reduced on average by 43% from 108 seconds to 62 seconds in the current app implementation. When analyzing a subset of the data on iOS devices, where image upload and processing is much faster, the time was reduced by 71% from 89 seconds to 26 seconds. Further gains are expected after redesigning the ad placement app.

In the old process of manual categorization, the user had to do 3.1 clicks on average to reach the desired leaf category, assuming that he knew the exact path. With the new categorization service, the click path was reduced to just 0.4 clicks on average.

Customer satisfaction with the new category suggestion service was very high, with 95% of the customers rating suggestions and the whole improved user experience as excellent or very good.

The Fashion Cam project received a lot of attention from the general public and computer vision community with the biggest success of winning the best poster award at the NVIDIA GTC Conference 2017 in Munich. And most importantly, it was a well-received feature by users' feedback.

5. Conclusions and future work

Both fine-grained classification and similarity search retrieval are difficult problems to solve, even more so with data that lacks additional annotations beside the basic single-label annotations. Still, as our projects have shown, it is possible to develop both accurate and fast services to the satisfaction of the end user.

Future improvements mostly lie in the further utilization of the textual data that accompanies each ad image. For some categories, e.g. services and jobs, ad titles provide more contextual information than the images themselves.

We are currently developing solutions to improve and expand the categorization service to inputs that combine both title and image, very similar to the recent advances presented in [20] and [21]. Another approach we are working on utilizes attention models for weakly supervised localization, similarly to ideas presented in [22].

To improve the similarity search service, besides the classification approaches, we are also preparing the necessary ground for similarity metric learning by using a triplet model, following [23]. Finally, we are also currently working on using user feedback to improve our similarity ranking.

References

[1] L. Deng, "Deep Learning: Methods and Applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3–4, pp. 197–387, 2014.

[2] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomputing, vol. 187, pp. 27–48, Apr. 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conf. on Computer Vision and Pattern Recognition, pp. 248–255, Jun 2009.

[5] T. Dettmers, Deep Learning in a Nutshell: Core Concepts. https://devblogs.nvidia.com/deep-learning-nutshell-core-concepts/ (March 2018.)

[6] M.-E. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes," in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp.722-729, Dec 2008.

[7] O.M. Parkhi, A. Vedaldi, A. Zisserman and C.V. Jawahar. "Cats and dogs." *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012): 3498-3505.

[8] A. Khosla, J. Nityananda, Y. Bangpeng and L. Fei-fei, "Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs." (2012).

[9] C. Wah, S. Branson, P. Welinder, P. Perona and S.J. Belongie. "The Caltech-UCSD Birds-200-2011 Dataset." (2011).

[10] J. Krause, M. Stark, J. Deng and L. Fei-Fei, "3D Object Representations for Fine-Grained Categorization," 2013 IEEE Int. Conf. on Comp. Vision Workshops, Sydney, NSW, 2013, pp. 554-561.

[11] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9.

[12] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017, pp. 6517-6525.

[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conf. on Comp. Vision and Pattern Recognition* (2017): 2261-2269.

[14] L. Deng, J. Krause, A.C. Berg and L. Fei-Fei. "Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition." *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012): 3450-3457.

[15] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[16] J. Wang, T. Zhang, j. song, N. Sebe and H. T. Shen, "A Survey on Learning to Hash," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 769-790, April 1 2018.

[17] H. F. Yang, K. Lin and C. S. Chen, "Supervised Learning of Semantics-Preserving Hash via Deep Convolutional Neural Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pp. 437-451, Feb. 1 2018.

[18] E. Hoffer and N. Ailon, "Deep Metric Learning Using Triplet Network," in Similarity-Based Pattern Recognition, Springer International Publishing, 2015, pp. 84–92.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A.Y. Ng. "Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011 (pp. 689-696)

[20] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, "Effective deep learning-based multi-modal retrieval," The VLDB Journal, vol. 25, no. 1, pp. 79–101, Jul. 2015.

[21] X. He and Y. Peng. "Fine-Grained Image Classification via Combining Vision and Language." *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2017): 7332-7340.

[22] X.He, Y. Peng and J. Zhao. "Fast Fine-grained Image Classification via Weakly Supervised Discriminative Localization." *CoRR* abs/1710.01168 (2017): n. pag.

[23] J. Wang et al., "Learning Fine-Grained Image Similarity with Deep Ranking," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1386-1393.