DOI: 10.3217/978-3-85125-954-4-15

Vision-Language Models for Filtering and Clustering Forensic Data

Axel Weissenfeld, Bernhard Strobl
AIT Austrian Institute of Technology, Center for Digital Safety & Security,
Giefinggasse 4, 1210 Vienna, Austria
axel.weissenfeld, bernhard.strobl @ait.ac.at

David Weichselbaum, Christopher Wimmer, Martina Tschapka
david.t.weichselbaum@gmail.com, cedwimmer@gmx.at, martinatschapka@gmail.com

Abstract

With image- and video-capable devices in the hands of a majority of the population worldwide, the amount of media data keeps growing. Hence, the search of specific images and clustering of datasets is of great importance to extract the relevant information, e.g. search for a specific person by legal enforcement agencies (LEAs). This paper presents a new tool which uses vision-language models to filter and cluster forensic data. The tool provides a GUI, which enables a flexible search by accepting textual as well as image input, to search large amounts of data in near real-time.

1. Introduction

The search for a specific person in images and videos is an important task in forensics and part of Content-Based Image Retrieval (CBIR) [3]. Unfortunately, a manual search is very time-consuming and a fully automatic search is usually not applicable. As a result, critical evidences might literally be hidden in plain sight, among an overwhelming number of images and videos.

In this work we do propose a system based on vision-language models assisting an operator to quickly filter and cluster image data by searching for pedestrian attributes. Pedestrian attributes are humanly searchable semantic descriptions such as gender, hair length, clothing style, or facial features and can be used as soft-biometrics in visual surveillance.

Pedestrian attributes recognition (PAR) is often approached as a metric learning problem [11], where one seeks to retrieve images containing the person attributes (Fig. 1). This is challenging in the sense that images captured by different cameras often contain significant intra-class variations caused by the changes in background, viewpoint, human pose, etc.

The developed solution shall enable the operator to filter and cluster image data by person attributes (Sec. 2). Some results are presented in Sec. 3.

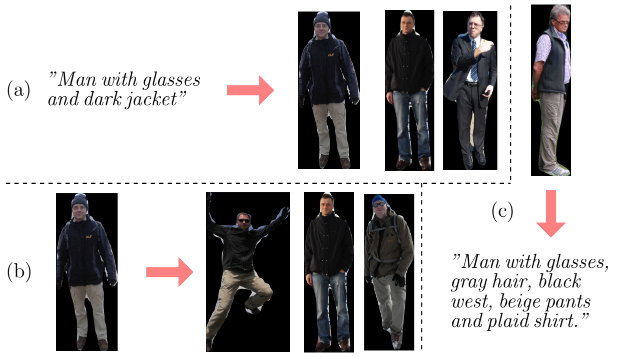
(a)
"Man with glasses and dark jacket"
(b)
(c)
"Man with glasses, gray hair, black west, beige pants and plaid shirt."

Figure 1. Vision-Language models enable a flexible search: (a) text→image retrieval, (b) image→image retrieval, (c) image→text retrieval

2. Vision-Language Models for Filtering and Clustering

More recently, CBIR systems have been extended by multimodal inputs such as image-text pairs, which we denote as Vision-Language (VL) models [2,6,10] . In contrast to prior models that are trained on images with class annotations, VL models are directly trained on image-text pairs to group relevant text vectors matching to the meaningful image content vectors. Recently some very large models such as CLIP [9], ALIGN [5], and BASIC [8] were trained, which achieve large robustness even on challenging datasets and a high accuracy with zero-shot classification. For instance, CLIP is a contrastive approach to learn image representations from text, with a learning objective which maximizes similarity of correct text-image pair embeddings.

VL models allow a textual search in image data by entering keywords or sentences. The possibility of extensive textual input is especially beneficial for PAR analysis, since different attributes can be combined. For example, if an operator searches for a man with a black backpack, the operator can enter ”*man with a black backpack*”. The model returns a confidence score (probability) that an image contains the searched attributes. The free text input simplifies the use of the system. There are also numerous application variations such as providing reference image as input or extracting semantic information from images (Fig. 1). Moreover, VL models also produce robust embeddings, which are indispensable to accurately cluster forensic datasets.

Optionally persons in the image data can be segmented (instance segmentation) and extracted from the original image. While object detection identifies objects in the image data, segmentation assigns an object class to each pixel. For the instance segmentation used here, a network architecture called Mask R-CNN is used [4]. The Mask R-CNN model was fine-tuned on the OpenImages[1] and Coco[2] datasets to segment persons. A single segmented person is the input to our VL model as illustrated in Fig. 1.

The developed tool for clustering and filtering image data enables LEAs to search their data in a targeted and focused manner, but also to conduct general screenings of large data sets (working with over 100.000 files) before a clear investigation target is defined. The VL model generates 512-dimensional embedding vectors of the image-text input. The tool automatically groups content into meaningful clusters using unsupervised machine learning [1] and arranges the input images by a nonlinear down projection. The tool also provides a simple GUI to search for specific persons by providing natural language search terms or an image or by selecting special trained classifiers. In addition to the reference text or reference image search, the third alternative for the analysis of image data is the implementation of particularly trained MLP (multi-layer perceptron) networks that were trained on several classes. By choosing a certain classifier inside the application, the provided image data is automatically classified. For each class, a distinct single-label classifier was trained.

3. Results

For generating the results we used the ViT-B/16 model [9]. Fig. 2 depicts the clustering result of a fraction of the PA-100k dataset as well as the search result for the reference image shown on the left. Similar results can be received if a textual description is the input.

The VL model reliably classifies pedestrian attributes as illustrated in Fig. 3. For illustration purposes the corresponding attention parameters are saved per residual attention block during the forward pass of an image and a backpropagation is conducted following the forward pass computation, with respect to the known output vector. After multiplying these two values (attention value and gradient) in the respective layers and the respective subspace, the computed information of all residual attention blocks is superimposed as a heatmap over the input image.

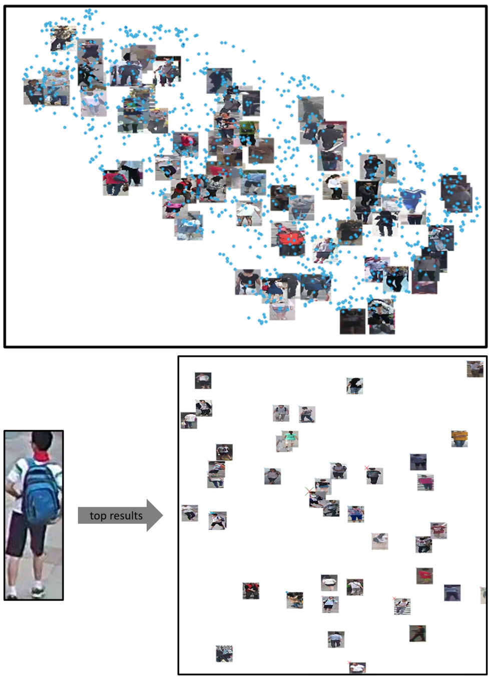

Figure 2. Top: Clustered and displayed result of a fraction of the PA-100k dataset. Bottom: Found top results for a sample image of the PA-100k dataset (provided by [7] under the CC BY 4.0 license[3]). Using a visual indication (red crosses) in the presented two-dimensional space, the software identifies the most relevant images to the user's particular search operation.

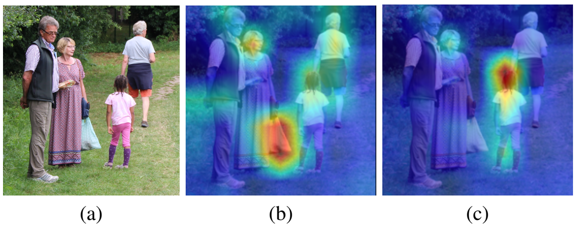

Figure 3. Searching in (a) for a bag and child. Resulting heatmaps are displayed in (b) and (c), respectively.

The filtering can be very efficiently executed and enables near real-time searches, since the embeddings are highly compact and only a dot product between text and image or image and image embeddings need to be carried out.

[1] https://opensource.org/licenses/MIT
[2] https://cocodataset.org
[3] https://creativecommons.org/licenses/by/4.0/

4. Conclusions

This work presents a new tool to search through forensic data. The tool is highly flexible because of the used VL models, which enable a clustering as well as search using pedestrian attributes or a reference image. Promising and encouraging results were obtained showing the feasibility of the tool for operational use by LEAs.

Note, that the solution is not limited to persons.

Acknowledgments

This research work has been supported by Vienna Business Agency.

References

[1] Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1(2665):2012, 2012.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[3] Venkat N Gudivada and Vijay V Raghavan. Content based image retrieval systems. *Computer*, 28(9):18–22, 1995.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[6] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[7] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017.

[8] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[10] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[11] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.