Knowledge Distillation for a Domain-Adaptive Visual Recommender System

Alessandro Abluton[1,2,*,†]

[1]Computer Science Department, University of Turin, Italy
[2]Inferendo srl, Alessandria, Italy

Abstract

In the last few years large-scale foundational models have shown remarkable performance in computer vision tasks. However, deploying such models in a production environment poses a significant challenge, because of their computational requirements. Furthermore, these models typically produce generic results and they often need some sort of external input. The concept of knowledge distillation provides a promising solution to this problem. By leveraging the teacher-student framework, the smaller "student" model learns to mimic the larger "teacher" model. In this paper, we focus on the challenges faced in the application of such techniques in the task of augmenting an object detection dataset used in a commercial Visual Recommender System that needs to detect items in various e-commerce websites, encompassing a wide range of product categories. We also present a simple solution to the problems we identified and propose a possible direction of future works.

Keywords

Knowledge Distillation, Object Detection, Computer Vision, Visual Search

1. Introduction

Visual Recommender Systems have emerged as a powerful tool in the field of e-commerce, providing personalized product recommendations based on visual similarity and user preferences. As described in [1], while traditional recommender systems primarily rely on user-item interactions, Visual Recommender Systems leverage image similarity and visual search techniques to enhance the recommendation process. The fundamental building blocks entailing these systems are:

- **image similarity and feature extraction**: at the core of a visual recommender or any Content Based Instance Retrieval system (CBIR) is the ability to quantify and compare visual content[2]. This involves extracting meaningful features from images, which can be extracted by means of statistical analysis such as a simple color histogram or can be produced by complex deep learning models as in the case we are studying.
- **visual search**: it is a crucial component of Visual Recommender Systems; it involves the detection of objects or items within user-uploaded images. This task is carried out by

AixiA 2023: 22nd International Conference of the Italian Association for Artificial Intelligence, Rome, Italy
*Corresponding author.
alessandro.abluton@unito.it (A. Abluton)
https://people.unito.it/persone/alessandro.abluton (A. Abluton)
0000–0001–8525–9940 (A. Abluton)

deep learning object detection models, enabling the system to identify products or items within the images with varying levels of accuracy. Most similar items to the one detected are then searched and retrieved.

In summary, to build a proper Visual Recommender System two models are needed:

1. **image embedding model**: it is responsible for producing embeddings of images that are capable of describing them in enough detail to perform a successful similarity search, e.g. embeddings of similar items must be as close as possible in the latent space.
2. **object detection model**: it must be able to recognize relevant items in images of products extracted from an e-commerce picture or from user-uploaded content.

While the use of a large pre-trained model such as CLIP[3] has empirically proven to be more than enough to produce accurate vector embeddings of the images, the dataset composition on which the object detection model is trained plays a pivotal role in determining the system's capability to recognize a broad array of objects. More importantly, the classes within the dataset correspond to the actual items within a potential e-commerce platform that would utilize the recommender system, making it a critical factor influencing the system's capacity to identify and classify diverse items.

In the present work, we aim at discussing a solution for a Visual Recommender System operating as a service, known as Recommendations as a Service (RaaS). Traditionally, recommendation systems were embedded within specific applications, requiring significant engineering effort and resources to implement and maintain. With the emergence of cloud computing and microservices architectures, the concept of RaaS has gained prominence.

The research project we are pursuing is being developed in the context of Visidea[4], a commercial product that aims to offer a wide range of recommender modalitites as a service, so that developers and businesses can add recommendation capabilities to their platforms through APIs, allowing seamless integration into various applications and websites. Visidea focuses on offering recommendations for the e-commerce sector and thus needs to be able to handle as many items as possible, due to the ever-varying nature of the markets.

While a Visual Recommender System offers significant benefits, it also faces a unique challenge known as domain adaptation. In this context, domain adaptation refers to the ability to accommodate a wide range of e-commerce websites that query the system, each one with its own set of products and classes of objects. To ensure accurate and relevant recommendations across different domains, the system needs to be able to quickly adapt and add new classes of objects to its object detection dataset.

E-commerce platforms, especially in the fashion industry, frequently introduce new clothing styles, accessories, or product categories to attract customers. For example, a new fashion e-commerce platform might start selling "Jumpsuits" a category of product that the existing object detection model may not be able to detect, because the original dataset on which it has been trained does not have that class in its labels. Traditional object detection models rely on extensive training data for each object class they are supposed to detect. When a new class appears, there is often a shortage of labeled training data, making it challenging to fine-tune the model effectively.

To tackle this issue, an automatic method is employed to add new classes to the object detection

dataset. This method leverages techniques like transfer learning and knowledge distillation [5] to efficiently transfer the knowledge from the large foundational models to smaller and faster models that can handle the new classes.

2. Knowledge Distillation through Autodistill

Distillation-based processes, similar to those used in natural language processing, have been applied to create more compact models that derive their knowledge from larger, pre-existing models. An illustrative instance of this approach is the Stanford Alpaca model, introduced in March 2023. The developers employed OpenAI's text-davinci-003 model to generate 52,000 instructions using an initial dataset as its reference point[6]. The Autodistill Python package developed by Roboflow company offers automated image labeling using foundational models that have undergone training on an extensive dataset, drawing on millions of images and significant computational resources invested by major industry leaders such as Meta, Google, and Amazon. The resultant dataset can subsequently serve as a valuable resource for training cutting-edge models, harnessing their efficiency and reliability for deployment in production environments. Autodistill provides a selection of base models and target models. One of the base models can generate a dataset in the precise format needed for the chosen target model. Consequently, training the target model on this generated dataset represents the actual "distillation" of knowledge

Amongst the pletora of base models offered by autodistill, we choose to employ GroundingDINO [7], an extension of the DINO [8] model developed precisely for the purpose of zero-shot object detection. DINO is a self-supervised learning method for visual representation learning. It introduces a novel training objective that encourages visual representations to emerge with consistent semantic properties. DINO achieves this by contrasting multiple views of the same image and optimizing a similarity-based loss function.

GroundingDINO is an extension of DINO that focuses on grounding visual representations with textual descriptions. It leverages the contrastive learning framework of DINO to learn representations that capture the semantics of both images and text. GroundingDINO performs joint training with image-text pairs and learns to align the visual and textual modalities by maximizing their similarity.

Autodistill needs as input both the images to label and a description, called "ontology", of the objects to detect inside those images. The ontology comes in the following format: `{"prompt": "label"}` where the prompt is a natural language description of the object to detect and the label is simply the associated class word. According to Autodistill an ontology can include several objects, each with its own prompt, that will be detected at the same time in each image. While the primary objective of autodistill is to facilitate the labeling of objects within images, an unexpected outcome was observed when dealing with ontologies containing more than one class. In these scenarios, the distillation process resulted in the labeling of all objects within the images with every possible class specified in the ontology. We identified two main issues:

- **label duplication**: several objects within the image were labeled with not only the appropriate class but also with almost every other class present in the ontology. Consequently, the correct label appeared multiple times, making the labeling output redundant.

- **erroneous labelling**: objects within the image were often mislabeled with classes that did not correspond to them. This meant that the distillation process was not only overly inclusive in assigning labels but also misinterpreted parts of the image as objects belonging to classes unrelated to the actual target.

These issues posed a significant obstacle in achieving precise and reliable object labeling, particularly when dealing with images containing multiple objects belonging to different classes.

Moreover, another important aspect is the sensitivity of the provided prompts. Even minor variations in the wording of prompts could have a profound impact on the quality and accuracy of the produced dataset. This sensitivity is notable when generating labels for specific objects, as it directly influences the model's ability to correctly identify and categorize objects within images.

To illustrate this issue, let's consider a practical example. Suppose the objective is to identify *swimsuits* within images. Initially, we provided the prompt "a picture of a swimsuit" to guide the model. However, this often resulted in the model not only correctly identifying the swimsuit itself but also erroneously associating the "swimsuit" class with the entire person wearing the swimsuit. In other words, the labeling extended beyond the target object to encompass the broader context.

Minor alterations of the prompt such as: "a single piece of a swimsuit." led to significantly improved results as shown in Figure 1. The model's ability to distinguish between the swimsuit as the target object and the person wearing it as the contextual background became notably more precise.

3. One-class-at-the-time solution

To take under control the issue related to the "multi-class identification" (i.e. the identification of objects not directly related to what one is actually searching for), a solution focusing on a single-class labeling was devised. This involves a one-class-at-a-time labeling strategy, where images of a single class, such as swimsuits, were collected from the web. The ontology provided for this approach contained solely the prompt and label for the specific class under consideration, omitting references to other objects. The prompt was manually created after some empirical testing on a subset of the downloaded images.

This one-class-at-a-time labeling approach yielded promising results, with the majority of images correctly labeled. The model demonstrated competence in identifying and associating the label with the intended class. Certain errors persisted, particularly when clothing is worn by humans. In such instances, the model occasionally struggled to differentiate between the garment, which was the intended target object, and the person wearing it, considered as part of the contextual background. This issue highlighted the complexities involved in recognizing objects within a contextual setting and remarks the need to find better methodologies to refine these foundational models, that frequently end up in being too generic to be actually used in a real world commercial application.

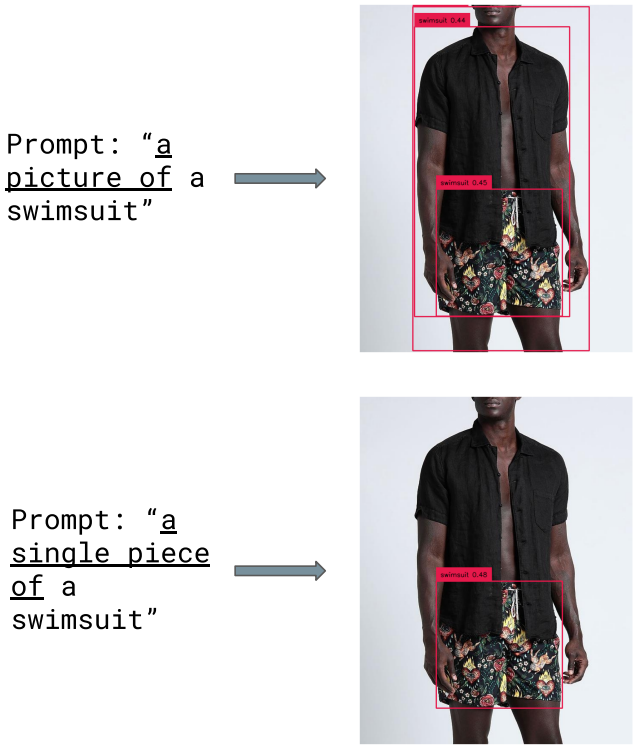

Figure 1: Example of the importance of choosing the right prompt, even if semantically the same, they yield completely different results. Image provided by Inferendo s.r.l

4. Conclusion and Future Works

Moving forward our research endeavors aim to overcome the challenges encountered in the single-class labeling approach within Autodistill. While this method has shown promising results in isolating and labeling objects, it still requires significant manual data collection efforts for each class and struggles in complex contextual scenarios. Our vision for the future involves the development of a more advanced and efficient solution that harnesses spatial and relational knowledge. This solution seeks to infer accurate object labels by analyzing the interplay between objects within an image. By leveraging sophisticated techniques for recognizing object relationships and spatial configurations, we aspire to enhance the quality and precision of image labeling in multi-class environments.

Acknowledgments

My reasearch is conducted as “Dottorato in Alto Apprendistato” in Inferendo, an innovative start-up, spin-off of the University of Piemonte Orientale. Funding as been provided by Regione Piemonte. I want to deeply thank my tutors Luigi Portinale (University of Piemonte Orientale) and Roberto Esposito (Univeristy of Torino) for their guidance and support throughout my

academic journey.

References

[1] A. Abluton, Visual recommendation and visual search for fashion e-commerce, in: International Conference on Similarity Search and Applications, Springer, 2022, pp. 299–304.

[2] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, M. S. Lew, Deep learning for instance retrieval: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 7270–7292. doi:10.1109/TPAMI.2022.3218591.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[4] Visidea, ???? URL: https://visidea.ai/.

[5] M. S. e. a. Gou J., Yu B., Knowledge distillation: A survey, International Journal of Computer Vision 129 (2021) 1789–1819. doi:10.1007/s11263-021-01453-z.

[6] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[7] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding dino: Marrying dino with grounded pre-training for open-set object detection, arXiv preprint arXiv:2303.05499 (2023).

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.