Knowledge Distillation for a Domain-Adaptive Visual Recommender System

Alessandro Abluton
Dipartimento di Informatica
Università di Torino,(Italy)
Inferendo srl, Alessandria (Italy)

Luigi Portinale
Computer Science Institute, DiSIT
Università del Piemonte Orientale, 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. In this paper, we focus on the challenges faced in the application of knowledge distillation techniques in the task of augmenting a dataset for object detection used in a commercial Visual Recommender System called VISIDEA; the goal consists in detecting items in various e-commerce websites, encompassing a wide range of custom product categories. We discuss a possible solution to problems such as label duplication, erroneous labeling and lack of robustness to prompting, by considering examples in the field of fashion apparel recommendation.

Introduction

Visually-aware Recommender Systems (VRS) have emerged as a powerful tool in the field of e-commerce, providing personalized product recommendations based on visual similarity and user preferences (McAuley et al. 2015; He and McAuley 2016; Packer, McAuley, and Ramisa 2018; Shankar et al. 2017; Tautkute et al. 2019). As described in (Abluton 2022), while traditional recommender systems primarily rely on user-item interactions, VRS leverage image similarity and visual search techniques to enhance the recommendation process. The fundamental building blocks entailing these systems are the following.

- **Image Similarity and Feature Extraction**. At the core of any content based instance retrieval system, such as a visual recommender, is the ability to characterize and compare different visual content (Chen et al. 2023). 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 supplied images, and a suitable retrieval strategy of similar objects. This task is carried out by 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 VRS two main modules are needed:

1. an **image embedding module**, responsible for producing embeddings of images that are capable of describing the images 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. an **object detection module**, 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 multimodal pre-trained model such as CLIP (Radford et al. 2021) has empirically proven to be more than enough to produce accurate vector embeddings of 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 VRS called VISIDEA[1] operating in a Recommendations as a Service (RaaS) setting. 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 micro-service architectures, the concept of RaaS has gained prominence; it arguably will transform the recommender system business landscape, since recommendation engines are an obvious domain for SaaS-based solutions, especially for SMEs.

[1] `https://visidea.ai`

A main obstacle to the deployment of an efficient VRS is the capability to deal with cross-domain applications (Kamani, Kumar, and Kagita 2023); this is particularly relevant when we want to build a recommendation platform able to recognize images from very different commercial categories such as fashion apparels, furniture or jewelry (which are all item categories where visual appearance is fundamental). In this context, the challenge of 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. Moreover, even in the case of a given specified domain (such as fashion apparels), also cross-catalog adaptability is very often necessary; this means that the VRS must be able to deal with different catalog categories from different e-commerce sites, each one using its own item characterization and classification.

To ensure accurate and relevant recommendations across different domains or catalogs, the system needs to be able to quickly adapt and add new classes of objects to its object detection dataset. As an example, in the fashion industry, e-commerce websites frequently introduce new clothing styles, accessories, or product categories to attract customers. A fashion e-commerce platform might start selling a category of products 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 effectively fine-tune the model. To tackle this issue, in this paper we propose tom exploit a knowledge distillation approach to map items contained in e-commerce images to the custom class labels of the e-commerce site. This method leverages techniques like transfer learning and knowledge distillation (Gou J. 2021) to efficiently transfer the knowledge from large foundational models to smaller and faster models that can handle the new classes. The main problem we address here is to fine-tune a visual model in order to adapt the target classes to a specific e-commerce site having a set of custom product categories. We focus on the fashion domain, where different e-stores may either sell different categories of clothing apparels or they may characterize the same category in different ways, for instance under different labels. As previously mentioned, the context we consider is that of VISIDEA VRS, an e-commerce platform plugin that aims to offer a wide range of recommender modalitites as a service, so that both developers and business operators can add recommendation capabilities to their platforms through APIs. It allows a seamless integration into various applications and websites, by focuses on offering recommendations for e-commerce sectors where image-based recommendation is crucial, and where the ever-changing nature of the markets is a factor.

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, preexisting models.

The *Autodistill* package[2] 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 (Liu et al. 2023), an extension of the DINO (Caron et al. 2021) 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 of the objects to detect inside those images which is called the “ontology”. 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 name. In particular, we restrict our attention to the so called `CaptionOntology` which is used to prompt a base model with a text caption and maps them into class names. The goal is to recognize specific prompted objects into the image and associate them to the given class label. In our setting, the aim is to map specific products pictured in a given image into the set of classes the VRS has to deal with.

According to *Autodistill*, a base model can be prompted in order to recognize and label several objects inside an image at the same time, (by specifying an ontology with several prompts, one for each object to be labeled). However, even if this facility appears to be functional to an efficient labeling of the objects within an image, we observed some unexpected results when dealing with ontologies containing more than one class label. In these scenarios, the distillation process resulted in the labeling of all objects within the images with every possible class specified in the ontology. In particular, we identified two main issues:

- **label duplication**: several objects within the image were

[2] `https://github.com/autodistill`

labeled with the expected label, but also with almost every other class present in the ontology.

- **erroneous labelling**: objects within the image were often mislabeled with classes that did not correspond to the expected one.

This means that the distillation process was both overly inclusive in assigning labels and also misinterpreting 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. If we provide a simple prompt such as *"a picture of a swimsuit"* to guide the model, we noticed that this may often result in the model not only correctly identifying the swimsuit itself but also erroneously associating the "swimsuit" class with the whole person wearing it (see Figure 1). In other words, the labeling extended beyond the target object to encompass the broader context.

Minor alterations of the prompt such as *"a swimsuit"* led to significantly improve the 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.

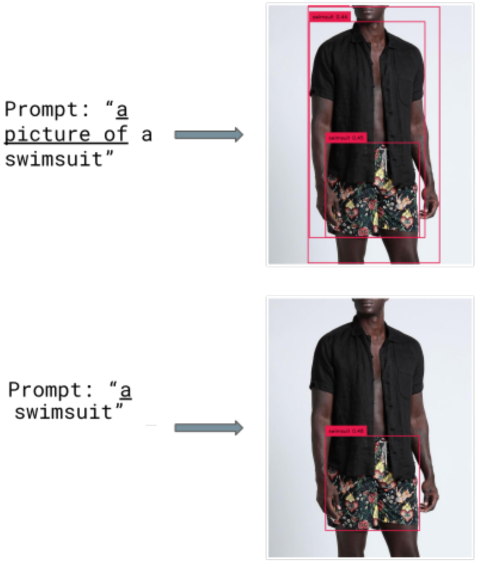

Figure 1: Example of the importance of choosing the right prompt

Class	Prompt	Prec	F1
pants	"pants"	0.856	0.922
pants	"trousers"	0.874	0.933
pants	"slacks"	0.870	0.930
pants	"full length trousers"	0.912	0.954
pants	"full length slacks"	0.906	0.951
sweater	"sweater"	0.411	0.582
sweater	"sweatshirt"	0.432	0.603
sweater	"jumper"	0.090	0.163
sweater	"pullover"	0.230	0.374
cardigan	"cardigan"	0.533	0.696
cardigan	"sweater"	0.800	0.889
cardigan	"sweatshirt"	0.706	0.827
cardigan	"jumper"	0.316	0.480
cardigan	"pullover"	0.500	0.667
jacket	"jacket"	0.806	0.893
jacket	"denim jacket"	0.731	0.844
jacket	"leather jacket"	0.845	0.916
jacket	"bomber jacket"	0.862	0.926
jacket	"parka"	0.063	0.118
jacket	"coat"	0.814	0.897
jacket	"trench coat"	0.644	0.783
jacket	"rain coat"	0.811	0.896

Table 1: Prompt evaluation: synonyms

In order to have a better understanding the above mentioned issues (i.e. the identification of objects not directly related to what one is actually searching for), we performed some evaluations. In particular, we considered a single-label prompting, where only one specific class is searched in the image. This resulted in a great mitigation of the errors related to label duplication since only one class is searched.

We selected a reference dataset among the ones used by VISIDEA containing 998 pictures of fashion apparels worn by people, distributed among 28 classes. We produced more than 5800 annotations on which to prompt the images. We have then performed several experiments by prompting the dataset in order to search specific items to map into the desired classes. We evaluated precision, recall and F1-score of the results. Firstly, we decided to test the sensitivity of the results to the prompt, by considering different variations of the prompt sentence (such as in the example provided in Figure 1), and we noticed that in general, specific words related to the searched items (e.g., "pants" or "full-lenght pants") were better that more complex prompting sentences (e.g., "identify the pants worn by the person"). This evaluation is particularly related to the precision dimension, since in all the experiments we obtained a perfect recall (i.e., no prompt missed the searched item, but several prompts returned items not related to the one of interest).

Secondly, we investigated the influence of synonyms on the results; this is particularly important in this context, since once we have decided to focus on a specific compact prompt (usually just a single word) to extract the item to map in the corresponding class label, then the problem is to decide the "right" word to use to get the desired result. Table 1 reports an excerpt of the evaluation on possible synonyms for some classes contained in the dataset. Since recall is always perfect (as already mentioned) we show here precision and F1-

score obtained on different target classes and prompts. We can notice that there may be situations where the identification process is rather robust to the change of the prompt (e.g., the "pants" class), while there are situations where the identification largely depends on the "right" synonym. In particular, it is interesting to notice the behavior on the "cardigan" class, which is one of the 28 reference classes in our dataset. The use of the exact class label is in this case significantly less precise that the use of a more general synonym such as "sweater". This is probably due to the fact that the distillation process has not been sufficiently trained on this specific class of product.

In conclusion, in order to face the problem of mapping visual information of items into the desired class, these results suggest a strategy where:

1. only a single label is searched in the image to be processed, eventually re-using the same picture to search for multiple items;
2. prompts are built by focusing on specific words that can be considered as reasonable synonyms of the class to be mapped.

The use of compact (single word) prompts is useful in the context of VRS where pictures of items usually contain a specific product to be put on sale; moreover, the adoption of a suitable set of synonyms allows one to be able to identify the right item to map in the custom class, even if the distillation process does not deal very well with the corresponding label (see the "cardigan" class example above).

Conclusion and Future Works

In the context of the VISIDEA VRS, the single-label prompting 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 human models are a relevant part of the picture. In such instances, the model occasionally struggled to differentiate between the clothing, 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 foundational models, that frequently end up in being too generic to be actually used in a real world commercial application.

In moving forward, our research endeavors aim to overcome the challenges encountered in the single-label 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 such as attention mechanisms or graph neural networks, we aspire to enhance the quality and precision of image labeling even in a multi-label prompt setting. Furthermore, we aim to evaluate the proposed solution on a diverse range of specialized tasks to assess its generalization capabilities and robustness. This will involve conducting extensive experiments on various datasets and comparing the performance of the proposed approach to existing state-of-the-art models. We believe that this research direction has the potential to significantly improve the performance of models in specialized domains and pave the way for more accurate and reliable automatic labelling systems.

Acknowledgements

A. Abluton's research is funded by Inferendo srl, an innovative start-up, spin-off of University of Piemonte Orientale, and by Regione Piemonte (Italy) under an advanced apprenticeship PhD scholarship.

References

Abluton, A. 2022. Visual recommendation and visual search for fashion e-commerce. In *International Conference on Similarity Search and Applications*, 299–304. Springer.

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

Chen, W.; Liu, Y.; Wang, W.; Bakker, E. M.; Georgiou, T.; Fieguth, P.; Liu, L.; and Lew, M. S. 2023. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(6):7270–7292.

Gou J., Yu B., M. S. e. a. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129:1789–1819.

He, R., and McAuley, J. 2016. VBPR: Visual bayesian personalized ranking from implicit feedback. In *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*.

Kamani, R.; Kumar, V.; and Kagita, V. 2023. Cross-domain recommender systems via multimodal domain adaptation. *arXiv preprint arXiv:2306.13887*.

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

McAuley, J.; Targett, C.; Shi, Q.; and van den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 43–52.

OpenAI. 2021. Openai gpt-3 api [text-davinci-003].

Packer, C.; McAuley, J.; and Ramisa, A. 2018. Visually-aware personalized recommendation using interpretable image representations. *arXiv preprint arXiv:1806.09820*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shankar, D.; Narumanchi, S.; Ananya, H.; Kompalli, P.; and Chaudhury, K. 2017. Deep learning based large scale visual recommendation and search for e-commerce. *arXiv preprint arXiv:1703.02344*.

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

Tautkute, I.; Trzciński, T.; Skorupa, A. P.; Brocki, Ł.; and Marasek, K. 2019. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access* 7:84613–84628.