Boosting image retrieval with accurately labeled small data

Amanuel Negash Mersha
mersha.aman@gmail.com

Addis Ababa Institute Technology

Abstract. Image retrieval is an application of computer vision that, given an image, a system returns a set of images with similar content. Through the advancement of deep learning in recent years, image retrieval has gotten a significant boost in performance. However, such success is apparent through training deep models using massive amounts of data, incurring significant costs on compute and labeling. Several works investigated if feature vectors extracted from deep models can be used for image retrieval tasks, while others fine-tuned the deep models for better performance. In this work, we examine if detailed annotation of small sets of images can be used to induce more semantic information into the feature vectors through a simple shallow network with just 3k training data. We show that this simple technique can improve the richness of feature vectors extracted both from shallow models such as PCA and deep models such as ResNet and Vision Transformer.

Keywords: image retrieval, deep learning

1 Introduction

Image retrieval is a task where given an image, a system returns a set of images that are similar to the given image based on a ranking algorithm. Such solutions are employed in search engines to find similar images given an image from a user. Google, Amazon, and Pinterest heavily use these systems to advance their users' experiences. It is also used for person re-identification purposes such as security systems and police departments. Overall, image retrieval systems rely on building an image descriptor model that outputs a vector that can represent the content of an image. [1] have employed principal component and latent discriminant analysis to generate the image descriptor. [2] evaluated different low-level image descriptors such as shape context, PCA-SIFT, and SIFT. They showed that SIFT-based descriptors perform better and their own extended version of SIFT outperforms other methods. However, these techniques are shallow, and other than low-level features, the descriptors do not have a rich feature set. Several deep learning-based solutions were proposed to generate a deep feature set that describes images well. [3] proposed an auto-encoder model where the middle vector representation is taken as a descriptor of an image. They converted that representation to a binary representation once more as it has many advantages

over the real-valued vector representation. Finally, this binary representation is used to query similar images from an array of other images. [4] proposed a stacked convolution neural network architecture where the model employs a separate stack of CNNs, max pooling, and local normalization layers that detect low-level local features. In their scheme, they developed an algorithm to collect similar images in a fine-grained setting and trained their model using the triplet loss function by providing a pair of similar and a pair of dissimilar images. More recently, [5] used a transformer model called Re-ranking Transformer (RRT) that merges local and global features by providing the transformer with pairs of images along with descriptor features extracted from the pairs. As deep learning relies on big data, these works rely on collecting massive datasets either manually or in semi-automated ways. Generating deep features without the need for a massive dataset has not been tried and this research focuses on that front.

2 Related Works

Soon after the success of the AlexNet [6] model in computer vision task, [7] evaluated the model for image retrieval. The authors showed that image descriptors extracted from AlexNet, both before and after fine-tuning, perform very well compared to the then state-of-the-art techniques such as Fisher's Vector [8]. In their experiments, fine-tuning the model with the target dataset improves the image retrieval performance significantly. However, fine-tuning still needs a significant amount of labeled image data.

In later work, [9] proposed a method to generate data using a 3D image extraction method using the new data to fine-tune a pre-trained model to improve performance in image retrieval tasks. Their method of generating hard positive and negative images allows one to train a model with a triplet loss, enabling the model to identify important similarities between images and use those features to query similar images. While this mechanism mitigates the data scarcity problem, fine-tuning the model with new data still incurs significant compute costs.

More recently, [10] has investigated the performance of features extracted from a pre-trained Vision Transformer (ViT) model introduced by [11]. The model achieved state-of-the-art performance in computer vision tasks after being pre-trained with ImageNet dataset [12]. [10] showed that the features extracted from the second to last layer have encoded representations of the images. Without fine-tuning the model, they compared the result with sophisticated models and the result was remarkable as it was either equal or very close to the state-of-the-art performance when evaluated on four benchmarks.

In this work, we propose a simple shallow-feed neural network that induces more information into the feature vectors extracted from different models. With just 3k training accurately annotated dataset, we aim to increase the image retrieval accuracy of the image descriptor (feature vectors).

Model	Direct	FCNN Trained		
		Task 1	Task 2	Task 3
PCA	5.99	6.09	6.23	6.71
Resnet18	10.34	13.05	11.56	12.08
Resnet34	9.9	12.04	10.71	11.85
Resnet50	10.04	13.26	11.59	12.55
Resnet101	11.05	14.9	15.24	13.39
ViT-B16	12.08	**16.24**	13.71	14.78

Table 1. Performance report of the tried models. The upper part of the table shows the performance of the image descriptors on the image retrieval task. The bottom part of the report shows the performances of the model after being trained on a fully connected neural network (FCNN)

3 Dataset

The growing e-commerce industry allows one to easily scrap large datasets and research. These datasets contain professionally shot, high-resolution images with rich annotations on them. [13] published such a kind dataset on Kaggkle.com. The dataset contains about 44k images with their corresponding labels. The labels are two kinds of. Firstly, every product is labeled with categorical features such as Gender (Man, Female), Color (Blue, Black..), Master Category (Apparel, Accessories), Sub Category (Topwear, Shoes, Bottomwear...), Article Type (Tshirt, Shirts..), Season (Summer, Fall...) and Usage (Casual, Sport). Secondly, each product is labeled with tags (words) that describe the product in a variety and arbitrary ways.

To train and test our proposed model, we sampled 3000 images for testing and 1000 for testing. The process of sampling the images follows the following procedure: 1) group images based on the full similarity of the categorical feature set. 2) Drop the groups that have below 40 images. 3) Randomly select 100 groups. 4) Randomly sample 40 images from each of the groups. At this point, the entire sampled images will be 4000, which was split into 3000 and 1000 for training and testing. This process guarantees that, as the task is a fine-grained image similarity ranking, the model gets enough images to learn fine-grain and intricate features and differentiate between similar images.

4 Methodology

Our technique starts with extracting image descriptors using some mechanism such as using PCA or extraction of features from deep models. These image descriptors are fed to a single feed-forward neural network layer for a transformation. This transformed representation is fed into a set of classifiers that predict different features of the image content and to the recurrent neural network that generates tags related to the image content. Figure 1 shows the schematics of the model.

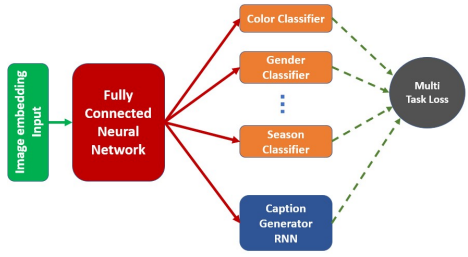

Fig. 1. The proposed model: The green block is a vector of an image extracted from either PCA or a deep model. The red block is a simple feed-forward network that transforms the vector. The Orange block represents classifiers for the Categorical Feature set of an image. Each block classifies a feature of the images such as color and season. The blue block is an RNN that generates the tag labels of the image. Finally, all the classifiers and the tag generator are trained will multi-task loss as shown by the black circular component.

Fig. 2. Comparing pre-trained Vision Transformer descriptor to fine-tuned descriptor. The first image of the first row is the query image. The first row presents a list of similar images based on ground truth. The second row is a list of images queried based on a pre-trained ViT descriptor. The third row is a list of images queried based on fine-tuned ViT descriptors. All three lists are sorted in that the similarity to the query image lowers as it goes from left to right. The numbers underneath each row represent the index of the true place of the image when sorted based on similarity. – shows that the image doesn't exist in the top 10 results of the ground truth presented in the first row.

The model is composed of a fully connected neural network (FCNN) to serve as a feature transformer. The latter part is the inducer, which induces information from the labels. Partly, this inducer is a categorical feature classifier. The tag of the images will be predicted using a GRU RNN. These modules can either be trained together or alone. We set the weights of the losses 1 so that all losses can be treated equally. The FCNN size is determined by the input vector. Our PCA feature was 2048, ViT feature is 768, Resnet18/34 is 512 and ResNet50/101 is 2048. The tag generator GRU has a hidden size of 512. Its input is also determined by the input feature size. The total unique tags were 292 after cleaning and normalizing. Hence the classifier on the top of the GRU has 292 sizes. In this way, just using highly accurate labels of the image, the model can incorporate new rich information into the original vector representation by transforming to a new feature space.

5 Experiment & Result

In the experiment, we used mean average precision (mAP) to measure the accuracy of the models. Mean average precision measures how two sorted lists are similar to each other by considering the items' location and a scoring function. PCA features of the images were used to create a baseline evaluation. The PCA feature size is 2048. We also evaluated 512 and 1024. However, their performance was very low and ignored from specifying in our report. Furthermore, image descriptors from each of the models were extracted to evaluate their performance on the dataset. Resnet18, Resnet34, Resnet50, Resnet101 [14] and ViT-B16 [11] models were to be used to extract the feature vectors.

In total, four kinds of experiments were done on each of the descriptor types. First, each feature vector is evaluated on the task without any modification. Secondly, the model is trained only on the categorical feature set of the image. Thirdly, the models were trained with the task of caption generation. Finally, all models were trained with the combination of the two tasks. Table 1 shows the performance of the models in each case.

Each experiment is run for a maximum of 200 epochs or stopped as soon as the loss becomes below 0.002. We used the Adam optimizer with a weight decay of 0.00001. While the weights of the models are initialized with Kaiming He initializer, the bias is set to 0.01 [15].

Overall, as the models go bigger and deeper, the representation quality of the vectors increases. In fact, as the number of layers of Resent increases, its performance on the ImageNet task increases. Once each vector is trained with our proposed method, its representation quality increases. Training solely on Task 1 (Categorical Feature set) shows the greatest advantage in all cases. Training with Task 2 (Tag generation) also improves performance but not as much as Task 1. Training on both tasks (Task 3) also improves but it fails to perform as much as Task 1. Table 1 shows that each model benefits from our method of fine-tuning with fine-grained labels. On average, each model improves by about

33% when compared to the original off-the-shelf descriptor after being trained on Task 1.

This phenomenon could be happening because of two reasons. As the objective function is a multitask objective, improper weights of the losses may affect and hence, may drag down the model's capability. In our work, we used equal weight for all the losses due to time constrain. Another explanation is the tags may not provide significantly important information for the model.

Figure 2 shows a sample result. It compares the prediction of off-the-shelf pre-trained Vision Transformer descriptors and fine-tuned descriptors with our method. The first image of the first list is the query image. It has the following labels: *women, apparel, topwear, tops, white, summer, casual.* That list of images is queried using the ground truth similarity metric. The second two lists are based on the shelf ViT and our fine-tuning method. The fine-tuned list contains mostly correct images even if the order is wrong. On the other hand, the ViT result contains mostly wrong images. Furthermore, the fine-tuned list contains only female models while the off-the-shelf ViT contains male models, which is wrong. This shows that our training method in fact induced the concept of female in the content of the original feature vector.

6 Conclusion

In this work, we investigated if additional accurate labels can be used to induce important information into feature vectors. We showed that such a technique can be used to fine-grained image retrieval tasks to boost the performance of the feature vectors. We evaluated different versions of the ResNet model, Vision Transformer, and a simple shallow PCA feature extractor. In future work, we plan to investigate the disparity created between the categorical feature sets and tag generation under the multi-task loss setting. The code for this work can be accessed from: https://github.com/leobitz/boosting_image_retrieval.git

References

1. Swets, D.L., Weng, J.J.: Using discriminant eigenfeatures for image retrieval. IEEE Transactions on pattern analysis and machine intelligence **18**(8) (1996) 831–836
2. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE transactions on pattern analysis and machine intelligence **27**(10) (2005) 1615–1630
3. Krizhevsky, A., Hinton, G.E.: Using very deep autoencoders for content-based image retrieval. In: ESANN. (2011)
4. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 1386–1393
5. Tan, F., Yuan, J., Ordonez, V.: Instance-level image retrieval using reranking transformers. arXiv preprint arXiv:2103.12236 (2021)
6. Alex, K., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional networks. In: volume-1; pages-1097–1105; NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems. (2013)

7. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European conference on computer vision, Springer (2014) 584–599
8. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: European conference on computer vision, Springer (2010) 143–156
9. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7) (2018) 1655–1668
10. Gkelios, S., Boutalis, Y., Chatzichristofis, S.A.: Investigating the vision transformer model for image retrieval tasks. arXiv preprint arXiv:2101.03771 (2021)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. (2009) 248–255
13. Aggarwal, P.: Fashion product images dataset (2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. (2015) 1026–1034