Augmenting Backbone Approaches To Instance-level Recognition For Artworks

Navneeth Krishna M.
nm3932@nyu.edu

Abstract

To tackle a dataset for large-scale instance-level recognition in the domain of artworks, we study broad domains of Convolutional Neural Networks as backbone structures of learning elemental 'art' class features. In addition, we expand the learning paradigm with augmented approaches in Instance Level Recognition (ILR) as visual recognition tasks on specific object instances. In our contribution, we perform Artwork Recognition (to recognize artworks in the images) using non-parametric approaches that attempt to etch profound accuracies on standard tests and compete with existing archetypes. The current benchmark exhibits a number of different challenges such as large inter-class similarity, long-tail distribution, and many classes. We work on the open-access collection of the MET museum art collection which is a large dataset of about 224K classes where each class corresponds to a museum exhibit with photos taken under studio conditions. Testing is primarily performed on photos taken by museum guests depicting exhibits in order to introduce a distribution shift between practical and ideal samples for emulating a real-world performance. A number of suitable approaches are evaluated to offer a testbed for future comparisons. Semi-supervised and Semi-weakly supervised models are effectively utilized to train the backbone which is used for non-parametric classification that is shown as a promising direction.

1 Introduction

Classification of objects into categories is in practice defined with different levels of granularity [4]–in demand of more distinct features within the class and to shy away from computational expenses. As aficionados and museums alike are striving for more precise image and artwork categorization, the reconnaissance of instance-level recognition has emerged at the forefront of potential gateways in addition to Contrastive Learning [5]. Instance-level recognition (ILR) is the computer vision task of recognizing a specific instance of an object. For example, instead of labeling an image as a "post-impressionist painting", we're interested in instance-level labels like "Starry Night Over the Rhone by Vincent van Gogh", or "Arc de Triomphe de l'Étoile, Paris, France", instead of simply "arch" [1]. ILR is applied to a variety of domains such as products, landmarks, and urban locations and has applications in visual search apps, personal photo organization, shopping, and more. Several factors make ILR a challenging task. It is typically required to deal with a large category set, whose size reaches the order of 106, with many classes represented by only a few or a single example, while the small between class variability further increases the challenge.

Despite its many real-world applications and encouraging aspects of the task, ILR has attracted less attention than category-level recognition (CLR) tasks, which are accompanied by large and popular benchmarks, such as ImageNet, that serve as a testbed even for approaches applicable beyond classification tasks. A major cause for this is the lack of large-scale datasets. Creating datasets with accurate ground truth at a large scale for ILR is a tedious process, As a consequence, many datasets include noise in their labels.

In this work, we focus on ILR for artworks and present findings from experiments conducted on non-parametric classification principles of deep learning. Our training dataset [3] is a collection of images from the Metropolitan Museum of Art (The Met) in New York. The training set consists of about 400k images from more than 224k classes, with artworks of worldwide geographic coverage and chronological periods dating back to the Paleolithic period. Each museum exhibit corresponds to unique artwork and defines its own class. The training set exhibits a long-tail distribution with more than half of the classes represented by a single image–making it a special case of few-shot learning.

2 Dataset

We use the Met dataset [3] for our experiments and we elaborate below on the constituents of the dataset.

2.1 Goal

The goal is to recognize the Met exhibits depicted in the Met queries. There is a distribution shift between these queries and the training images which are created in studio-like conditions. It includes a large set of images that aren't related to The Met, which forms an Out-Of-Distribution (OOD) query set, the distractor queries. There is no Met exhibit depicted in distractor queries. The query set is composed of the combination of those two sets.

2.2 Scale and splits

This dataset for ILR contains two types of images: exhibit images that are from the open-access collection and query images from the public-contributed sources. The Met collection was taken by the Met organization under studio conditions while capturing multiple views of objects featured in the exhibits. These images from the training set for classification are interchangeably referred to as exhibit or training images. It contains about 397k exhibit images corresponding to about 224k unique exhibits, i.e. classes, also called Met classes. The exact breakdown is as observed in Figure 1.
The Met dataset (Figure 2) contains about 20k query images that are divided into the following three types: 1) Met queries, which are images taken at The Met museum by visitors and labeled with the exhibit depicted, 2) other-artwork queries, which are images of artworks from collections that do not belong to The Met, and 3) non-artwork queries, which are images that do not depict artworks. The last two types of queries are referred to as distractor queries and are labeled as "distractor" class which denotes out-of-distribution queries.

Split	Type	# Images			#Classes
		Met	other-art	non-art	
Train	Exhibit	397,121	-	-	224,408
Val	Query	129	1,168	868	111 + 1
Test	Query	1,003	10,352	7,964	734 + 1

Figure 1: The Data Classes and Number of Images

The data collection and annotation process are described as follows and also summarized:

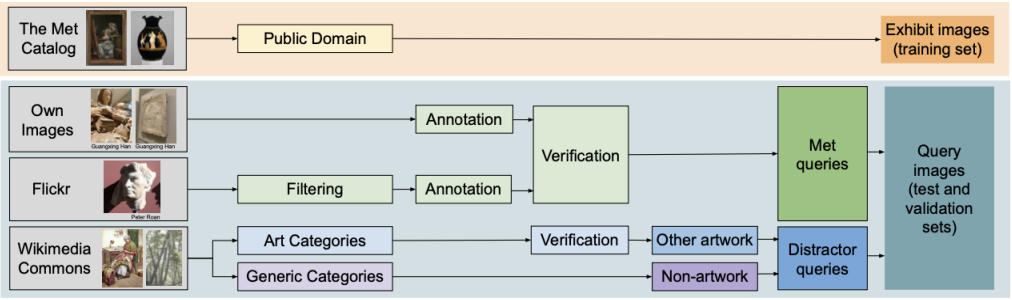

Figure 2: The Met dataset collection and annotation process

Because of the bulk of classes in the dataset, a workable subset with nearly 35000 classes is preferred for quicker understanding and descriptor extraction of the dataset. This segment is known to be 'The Mini MET' dataset and will be in use for similar purposes as the bulkier dataset.
The Met is the largest ILR dataset in terms of the number of classes that have been manually verified. Overall, the Met dataset proposes a large-scale challenge in a new domain, encouraging future research on generic ILR approaches that are applicable in a universal way to multiple domains.

2.3 Mini Dataset

The Mini Met dataset has over 39,000 images from the MET data collection and is a competitive mini version of the actual dataset. These images are randomly picked from different classes and are therefore not manually skewed to contribute to inter-class similarity. It contains the same classes as the original and has been collected for experimentation use to study the variation in benchmark results and for the convenience of extractor description. We have used this dataset in our experimentation procedure and we highlight the marked performance of this dataset. A sample of its classes are as seen in Figure 3 and some prediction highlights are observed in Figure 4.

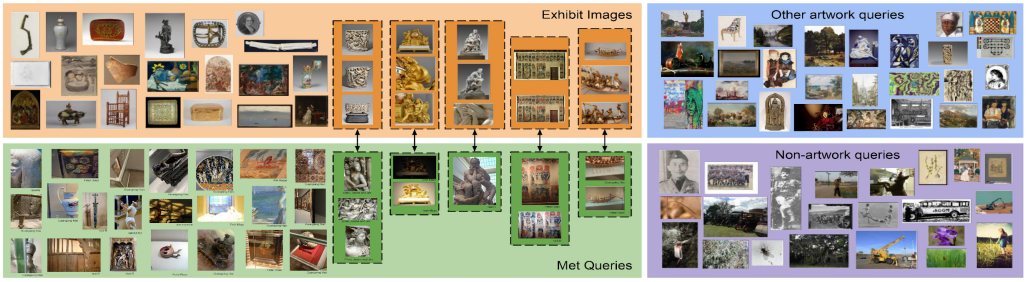

Figure 3: The images and classes in both Mini and MET dataset

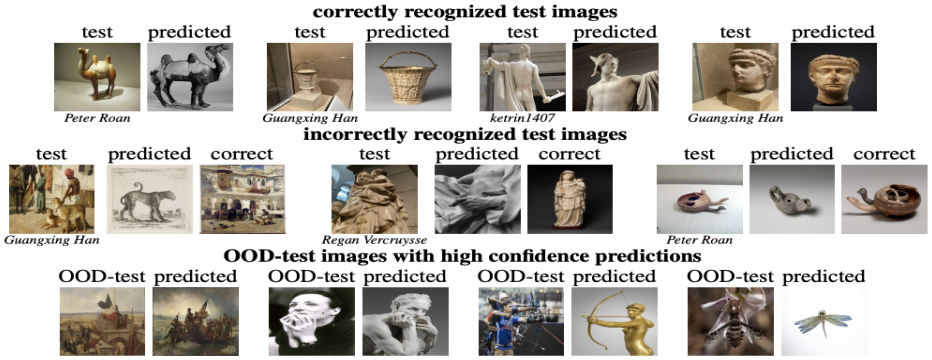

Figure 4: Images from MET dataset for high-performance approach. The correct predicted and incorrect predictions are shown.

3 Model

In this section, we outline the core idea that enabled the Model, its architecture, and the hyper-parameters that were tuned with Grid Search. Our goals experimented with limited scope in the following domains:

Parametric Classification. The Parametric Classification presents a Contrastive Learning approach to the descriptors through baseline techniques and concludes with classification. Our focus drifted away from the window of parametric classification as it was proven [3] that non-parametric classification outperformed its counterpart. Thus, our primary focus was constructed on the backbone models utilized in the non-parametric approach.

Non-Parametric Classification. We prioritize the non-parametric approach which consists of post-processing steps prior to classification with K-nearest Neighbors. The non-parametric classification approach initially deals with iterating through the MET dataset and performing Descriptor extraction using a ResNet model. This ResNet model that is chosen for the experiment is referred to as the ‘backbone’. Results from various backbone architectures are studied and have been presented. We use the standard Data Loader and torch vision libraries for the training, data loading, and deep learning processes.

Representation of Images. Let us consider an embedding function $f_\theta : \chi \rightarrow \mathbb{R}^d$ which takes an input image $x \in \chi$ and maps it to a vector $f_\theta(x) \in \mathbb{R}^d$ *also denoted by* $f(x)$. The function comprises a fully convolutional network as the backbone (parametrized by the parameter set θ), a global pooling method (shown to be effective for representation in instance-level tasks [12]) that maps a 3D tensor to a vector (using Generalized-Mean pooling [11]), a vector l_2 normalization, an optional fully-connected layer with 1 x 1 convolution as a projector, and a final vector l_2 normalization. Once the backbone processing is done, based on standard practice in instance-level search, the image representation space is whitened using PCA whitening [13] that is learned on the representation vectors of all the training images. We also perform dimensionality reduction as a final step. Representation of image x, denoted by vector embedding $v(x) \in \mathbb{R}^d$, is an aggregation of multi-resolution embeddings and is given by

$$v(x) = \frac{\sum_{r \in R} f(x_r)}{\left\| \sum_{r \in R} f(x_r) \right\|},$$

where x_r denotes image x that is down-sampled by a relative factor r. The value of R is constantly $\{1, 2^{-0.5}, 2^{-1}\}$ as we prefer a multi-scale representation.

3.1 Architecture

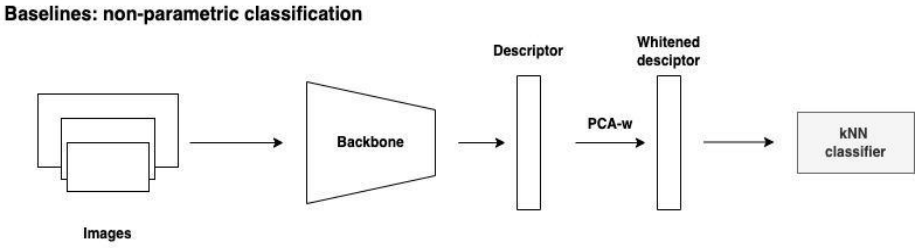

Figure 5: Non-parametric classification approach

In our approach, we rely on the 'backbone' component that embeds a non-pre-trained standard CNN model, from a set of alternatives (Figure 5). So far, we tried experimenting with other alternatives and explored possibilities related to advancing various accuracies. The models that are used as backbone are chosen as follows:
ImageNet (IN): Trained on ImageNet with Cross-entropy Loss; *StylizedImageNet (SIN)*: Trained on ImageNet by Geirhos et al. [14]; *SwAV on ImageNet (IN) - self supervision*: representation learning on ImageNet with self-supervision by instance discrimination. The resulting network has achieved good results in concept generalization [15]; and *Semi-weakly supervised (SWSL) on Instagram 1G + ImageNet*: teacher-student approach [10] with teacher pretrained on about 1 billion images with hashtags and student trained with teacher-generated pseudo-labels, eventually fine-tuned on ImageNet. We experiment with variants of Residual Neural Networks trained on the above backbones.

Once extracted, the descriptors from the backbone are "pickled" in and out of the backbone and into the kNN Classifier. The descriptors undergo PCA whitening in order to make the input less redundant. The kNN Classification occurs with an auto-tuning mechanism and it relies on the training, validation, and testing descriptors saved from the backbone phase. We tune the pooling methodology along with the sensitivity of the end of the backbone. Finally, the kNN evaluates and provides the GAP and ACC scores with and without distractors based on an auto-tuning mechanism that is presented in the next section.

3.2 Tuning of Hyper-parameters k and τ

Hyper-parameters k and τ have been tuned using grid search according to the GAP value on the validation set. For the kNN classifier, for an image x, its label is denoted by $y(x)$ and q is a query image. To find similarity between query and training image, we use $v(x)^{\top} v(q)$. The confidence of the class c for query q is given by

$$s_c(q) = \max_{x \in NN_k(q)} (v(x)^T v(q)) \mathbb{1}_{y(x)=c},$$

where $NN_k(q)$ is the set of k nearest-neighbors of q in the d-dimensional representation space and $s(q) \in \mathbb{R}^N$ is the vector of class confidences with elements $s_c(q)$ where N is the number of training classes. While the normalized confidence is given by taking soft-max of vector $\tau s(q)$, where τ is temperature. Although the label prediction requires only k = 1, confidence estimation for more classes is essential for normalization and handling out-of-distribution queries.

For each backbone architecture presented in 3.1, our kNN algorithm works on the train, validation, and test descriptors by fine-tuning and then chooses the most optimal set of k and τ that produces the GAP, GAP−, and ACC scores.

4 Results & Experimentation

The structure and evaluation protocol for this problem mimics that of the Google Landmarks Dataset [16]. For evaluation, we measure the classification performance with two standard ILR metrics, namely average classification accuracy (ACC), and Global Average Precision (GAP). The average classification accuracy is measured only on the Met queries, whereas the GAP, also known as Micro Average Precision (μAP), is measured on all queries taking into account both the predicted label and the prediction confidence. All queries are ranked according to the confidence of the prediction in descending order, and then average precision is estimated on this ranked list; predicted labels and ground-truth labels are used to infer the correctness of the prediction, while distractors are always considered to have incorrect predictions. GAP is given by

$$\frac{1}{M} \sum_{i=1}^{T} p(i)\, r(i),$$

where p(i) is the precision at the position i, r(i) is a binary indicator function denoting the correctness of prediction at position i, M is the number of the Met queries, and T is the total number of queries. The GAP score is equal to the area-under-the-curve of the precision-recall curve whilst jointly taking all queries into account. We measure this for the Met queries only, denoted by GAP−, and for all queries, denoted by GAP. In contrast to accuracy, this metric reflects the quality of the prediction confidence as a way to detect out-of-distribution (distractor) queries and incorrectly classified queries. It allows for the inclusion of distractor queries in the evaluation without the need for distractors in the learning; the classifier never predicts an "out-of-Met" (distractor) class. Optimal GAP requires, other than correct predictions for all Met queries, that all distractor queries get smaller prediction confidence than all the Met queries.

For all the backbone network variants, Table 1 and Table 2 provide the results of the above benchmarks for fine-tuned values of k and τ:

Model	k	τ	GAP	GAP−	ACC
R18IN [7]	3	50.0	27.85	48.01	52.71
R50IN [7]	7	50.0	31.58	51.77	55.81
R50SwAV [9]	15	50.0	41.73	56.49	58.91
R50SIN [8]	5	50.0	26.17	44.89	48.06
R18Sw-Sup [10]	5	50.0	35.92	56.15	59.68
R50-Sw-Sup [10]	5	50.0	49.09	69.00	71.31
ResNeXt-50-32x4d-SWSL [10]	5	50.0	26.07	50.20	53.48
ResNeXt-101-32x4d-SWSL [10]	5	50.0	25.38	49.72	52.71

ResNeXt-101-32x8d-SWSL [10]	50	50.0	20.13	39.19	44.18
ResNeXt-101-32x16d-SWSL [10]	7	100.0	17.85	42.49	47.28

Table 1: Benchmark Test Results for various Backbones on the Met dataset

Model	k	τ	GAP	GAP−	ACC
R18IN [7]	2	100.0	39.10	59.75	62.01
R50IN [7]	3	50.0	40.98	62.50	64.34
R50SWaV [9]	5	50.0	51.08	66.29	68.21
R50SIN [8]	3	50.0	40.81	57.75	60.46
R18Sw-Sup [10]	3	50.0	47.84	65.79	68.21
R50Sw-Sup [10]	15	50.0	60.00	74.95	75.96
ResNeXt-50-32x4d-SWSL [10]	3	50.0	40.22	60.15	62.01
ResNeXt-101-32x4d-SWSL [10]	3	50.0	37.50	62.47	64.34
ResNeXt-101-32x8d-SWSL [10]	3	50.0	32.87	55.95	60.46
ResNeXt-101-32x16d-SWSL [10]	3	50.0	25.63	56.50	59.68

Table 2: Benchmark Test Results for various Backbones on the 'Mini' Met dataset

Comparison with existing models. In the original work where the dataset is presented [3], our results confidently perform with higher accuracies. For instance, the R50IN model in the original approach produces GAP: 22.2, GAP−: 41.8, and ACC: 46.4 versus 31.58, 51.77, and 55.81 accuracies from our approach as observed in Figure 6.

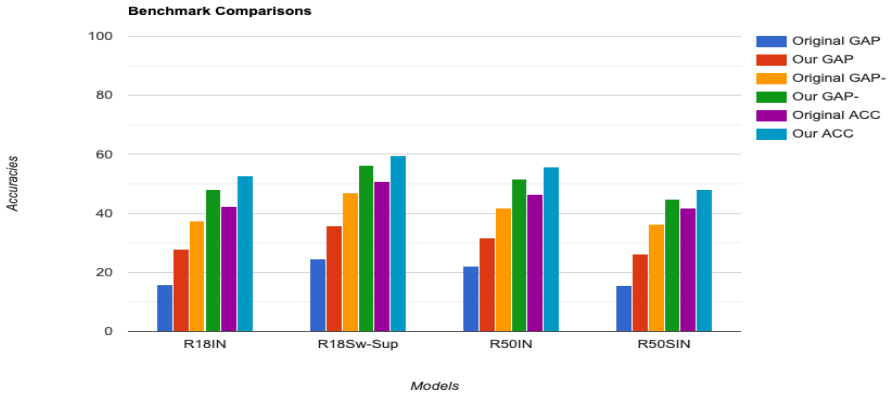

Figure 6: Benchmark Comparisons between Standard & Proposed works

An interesting phenomenon was observed in the increase in performance of the backbone when trained on the mini dataset compared to the entire dataset. Because of large inter-class similarity, the train descriptors on pretrained network backbones have been statistically more accurate in prediction when learning 39k training images instead of 390k. This leads us to the ever-challenging tradeoff between overfitting and underfitting and regulating model sizes. The results can be re-created by following the steps in the GitHub repository [2].

5 Conclusion

Across the destined timeline of this project, we have covered ground from learning the representational features of the dataset to uncovering challenges in the implementation roadmap. With such intuition, we cherish our exploration into the performance of such CNN approaches as the skeleton on this dataset via hyper-parameter

tuning, varying network backbones, and global pooling methods. We aim to further interpret the ins and outs of the deep learning tasks on these artistic grounds. Our results can be reproduced through our GitHub repository [2].

The auto-tuning of hyperparameters k and τ makes the process of determining the ideal parameter values more convenient for estimating the GAP and ACC scores. The discussion of single- versus multi-scale image representation yielded a lot of fruitful results in accuracies. While we could produce higher accuracies for the 3 important benchmarks compared to standard results of backbones, we aim to further develop the performance of the model through advanced techniques such as other global pooling techniques for ResNets, replacing the backbones with CNN & non-CNN architectures, and by training ambitious descriptors to study their performance on the benchmarks.

References

[1] Askew, Cam, and André Araujo. "Advancing Instance-Level Recognition Research." *Google AI Blog*, 25 September 2020, https://ai.googleblog.com/2020/09/advancing-instance-level-recognition.html.
[2] https://github.com/navneeeth/hpml-final-project
[3] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne van Noord, Giorgos Tolias (2021). The Met Dataset: Instance-level Recognition for Artworks. Part of Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 pre-proceedings (NeurIPS Datasets and Benchmarks 2021).
[4] Tolias, Giorgos and Jenicek, Tomas and Chum, Ondřej. Learning and aggregating deep local descriptors for instance-level recognition. arXiv:2007.13172
[5] Khosla, Prannay and Teterwak, Piotr and Wang, Chen and Sarna, Aaron and Tian, Yonglong and Isola, Phillip and Maschinot, Aaron and Liu, Ce and Krishnan, Dilip. Supervised Contrastive Learning. arXiv:2004.11362
[6] Jin, Ming and Zheng, Yizhen and Li, Yuan-Fang and Gong, Chen and Zhou, Chuan and Pan, Shirui. Multi-Scale Contrastive Siamese Networks for Self-Supervised Graph Representation Learning. arXiv:2105.05682
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In ICLR, 2019.
[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In NIPS, 2020
[10] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. In arXiv, 2019.
[11] Filip Radenovic, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. PAMI, 41(7):1655–1668, 2019.
[12] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In ECCV, 2020.
[13] Hervé Jégou and Ondřej Chum. Negative evidence and co-occurrences in image retrieval: The benefit of pca and whitening. In the European conference on computer vision. Springer, 2012.
[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shap bias improves accuracy and robustness. In ICLR, 2019.
[15] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In NIPS, 2020.
[16] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval. In CVPR, 2020.