Exploring Themes and Bias in Art using Machine Learning Image Analysis

Sudeepti Surapaneni
School of Data Science
University of Virginia
Charlottesville, USA
ss9ud@virginia.edu

Sana Syed
School of Data Science
University of Virginia
Charlottesville, USA
ss8xj@virginia.edu

Logan Yoonhyuk Lee
School of Data Science
University of Virginia
Charlottesville, USA
yl2vq@virginia.edu

***Abstract*—Machine learning and computer vision have been applied for image recognition of art objects such as paintings, sculpture images etc. In particular, deep learning methods for image classification in art have been used to improve user engagement by providing access to accurately labelled and classified art objects. As an increasing number of notable museums turn towards creating open access collections, alternatives to the use of laborious human annotating methods are needed. This paper focuses on the Open Access initiative of The Metropolitan Museum of Art (The Met) which was launched in 2017 in an effort to expand The Met's reach and presence. The museum now provides a select dataset of information on more than 470,000 artworks in its collection for unrestricted commercial and noncommercial use. However, with a widely accessible collection, the Met now faces the problem of how to enhance the user experience via access to accurately labelled art. This paper focuses on machine learning methods with applicability to automated classification of images obtained from The Met's online collection. We aimed to: 1) Compare three different convolutional neural networks - ResNet 50, ResNet 101, and Inception-ResNet-V2- using human annotated data, 2) Add transparency and interpretability to our models by using Gradient Weighted Class Activation Maps (GRAD-cams) and to explore bias in gender labels and 3) implement a multi-label classification model using ResNet 50. Future work would include the use of unsupervised clustering methods / auto-encoders to explore additional themes in the data. Other extensions of this work would include exploring methods to implement fine grained visual categorization, to mitigate bias, and to address the limitations associated with culture and stylistic interpretations. Deep learning techniques for art image classification may also help detect consistent features of bias in human annotated art.**

***Index Terms*—deep learning, image classification, CNNs**

I. Introduction

The Open Access art movement has gained momentum globally with The Metropolitan Museum of Art being one of the leaders of this initiatives [1]. With large volumes of online art, museums are now seeking to leverage machine learning to enhance their user experience. A critical challenge exists in creating accurate, well-curated labels/ classifications for online art objects. As more museums are increasingly opening access to large collections of digitized artworks, the field of computer vision is also responding to this shift with significant contributions. From this perspective, the use of machine learning image analysis is a natural solution to this annotation dilemma. Solutions found within computer vision can be harnessed and leveraged to contribute to this space. This will help make museums like the The Met, to serve as stronger conduits for knowledge proliferation and preservation. Deep learning techniques for machine learning/ computer vision such as Convolutional Neural Networks (CNNs) can learn complex features from human annotated art and help solve the annotation problem [2]. These trained CNNs can then be used to annotate un-labeled art thereby automating this process. We have focused this work on comparing three CNNs which have been extensively applied across many different image classification tasks. We also applied visualization methods to understand the "black box" of machine learning based CNN image analysis. Finally, since the current work depends on human annotations, we also used our black box visualization methods to explore themes of bias in our annotated data.

In the current digital age, many art museums such as The Met have chosen to share digital copies of their art online. In the case of The Met, given the sheer volume of their online objects - nearly half a million images online, they are facing new and unprecedented challenges. Most of these arise due to the unfettered access for users to experience the breadth and the depth of The Met collection. The Met seeks to give its users the ability to refine and filter works of art. This for example could be based on the subject of interest regardless of the era, medium, and time. In order to allow this level of attribute based filtering/ categorization, The Met is seeking to leverage machine learning based image classification. Their overall aim is to determine the best method/process for predictive fine grain attribute categorization of The Met Open Access collection. In this paper, we show preliminary results from three different Deep Learning image analysis platform for automated data labeling using art objects from Open Access Met collection.

A Convolutional Neural Network (CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other [2]. CNNs are capable of learning complex features and spatial information of images. However, the computation time of training the model increases with the size of the dataset. In order to increase our time efficiency, we decided to use neural networks which were pre-trained on ImageNet which

has more than a million high resolution images of 1000 different types of objects [2]. There are many different model architectures that use convolutional layers trained on the ImageNet database, but deeper networks do not necessarily improve model performance for image classification. In deep CNNs, degradation problems emerge because accuracy gets saturated and degrades quickly [3]. However, residual learning can avoid this problem by using a stochastic gradient descent to skip the connection. Thus, we implemented ResNet 50, ResNet 101, and Inception-Resnet-V2 which are based on the residual learning framework.

II. Related Work

As a result of previous research and publications, models were developed to predict a painting's style, genre, and artist by leveraging aesthetic-related semantic-level judgments[4]. There has been research conducted to extract and measure visual similarity between paintings and for the process of classifying artwork based on artist, genre and or era [5][6]. For these tasks that marry computer vision and art analysis, researches have explored and devised ways in which they can discover the low level visual features including brush strokes and color[7]. These low level features offer insight into genre, artistic style and or artistic movement to which the artwork belongs to[8]. State-of-the-art deep residual neural networks including ResNet50 were implemented and fine-tuned to accomplish the tasks described above.

III. Data

As outlined in the data pipeline in Figure 1, we extracted 469,301 high-resolution images in addition to all the related object metadata from The Met's API. Each object's metadata was stored as a JSON string consisting of detailed information such as object ID, origin, tag, and other pieces of artist information, in addition to its corresponding image URL link and 40 other features. We stored the metadata within a SQLite database which was easily integrated with Python. The web-scraped images were stored on Rivanna which is the high-performance computing cluster at UVA. There were a total 1,039 tags and most of the objects were tagged as 'men' and 'women' which shows the bias within the collections.

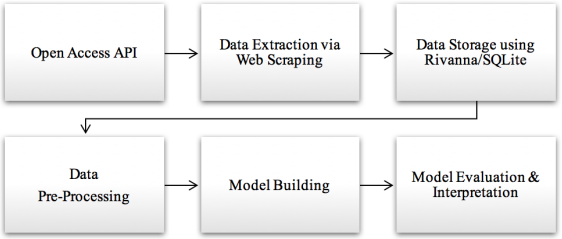

Fig. 1. Data Pipeline

A subset (n=233,000) of the image data we had access to from The MET had been labelled by human annotators with a total of 1,000 unique tags. Many of these images had multiple tags. First, for our single-label multi-class classification model we used the first tag listed. Since we wanted to have enough images for each tag, we filtered the tags to the top 23 tags. After filtering the data, we were able to use 97,010 images to train our CNN models. The data was split into 80 percent training, 10 percent validation, and 10 percent testing set. The data pipeline is shown if Figure 1, and the detailed data split is shown in Figure 2.

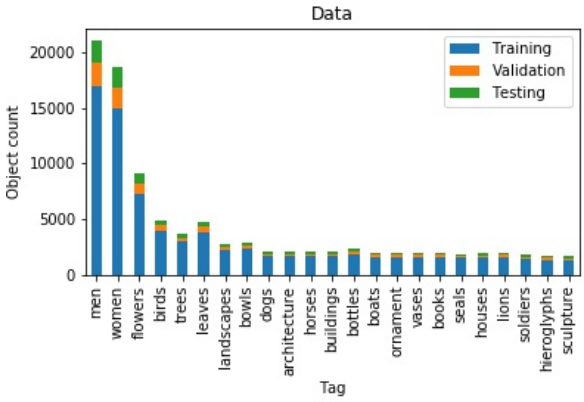

Fig. 2. Data Description

IV. Methods

A. Pre-Processing/Augmentation

To augment our dataset with images labelled as 'men' and 'women' we applied standard image transformations [9] to the training and validation sets. These include 40 degree random rotations, random horizontal flips, 0.2 width shift ranges, 0.2 height shift ranges, 0.2 shear ranges, 0.2 zoom ranges, and 10 channel shift ranges. The images were resized based on the model architecture. Images were resized 224x224 for ResNet 50 and ResNet 101. Images were resized 299x299 for Inception-ResNet-V2.

B. CNN Architecture

Three residual learning models' architectures pre-trained on ImageNet were compared. We added a single fully connected dense layer and a softmax layer at the end of the pre-trained models. During training the layers were unfrozen so that the models could update their weights. The models were trained on 25 epochs with batch size of 32. Adam was used as an optimizer with a learning rate of 1e-5. Categorical-cross entropy was used as a loss function, and accuracy was used as a metric.

ResNet 50, ResNet 101, and Inception-Resnet-V2 are three standard CNNs which have been used extensively in image classification tasks. The ResNet-50 model [2] consists of 5 stages each with a convolution and Identity block. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers. The ResNet-50 has over 23 million trainable parameters. The ResNet-101 shares almost

similar architecture in terms of the convolution block and identity block . However, ResNet-101 has a total of 101 layers and has over 42 million trainable parameters [2]. Inception-ResNet-V2 has the topological depth of 562 layers which includes all the activation layers, batch normalization layers and has about 55 million parameters[10]. Inception-Resnet-V2 is a hybrid given that it is considered to be a combination of the inception structure and the residual connections. Each inception-resnet block consists of multiple sized convolutional filters which are combined by residual connections. Each Inception-ResNet block is followed by a 1×1 convolution without activation. A batch-normalization layer is only used for the top of the traditional layers, but are not layered above the summations, further increasing the number of inception blocks used[10]. Inception-ResNet-V2 is considered to be a hybrid inception neural network with significantly improved performance in terms of top-1 and top-5 accuracy than its previous variants as well as resnet architectures[10]. The inception architecture that has been proven to perform very well at a relatively low computational cost. Due to the model's proven effectiveness, we decided to implement it as our third model.

C. Multi-Class Classification

A pre-trained ResNet50 was used to classify using 23 labels from the MET's dataset. During train time the following augmentations were performed randomly across the dataset: rotate, zoom, lighting, warp, affine. For classification we obtained probabilities for each class (via Softmax) using 0.2 as the threshold for classification. While training the model the accuracy and the f-score were used to prevent over-fitting. For a pretrained network where only the last linear layers were trained, a learning rate of 0.055 was used. Next, all the layers of the pre-trained network were allowed to train (un-frozen) using a linearly decreasing learning rate such that the initial layers have a smaller learning rate (10 to the power -3 versus the later layers 10 to the power -2). For the last layer a learning rate of 0.055 was used.

V. Results

A. Model Comparisons

Loss and accuracy of training and validation were obtained for three models after running them 25 epochs. Figure 3 shows that the training losses of three models decrease gradually over epochs, but the validation losses increase with some fluctuations. This indicates that the models are not learning from the data. The validation accuracy plot shows that the InceptionResnet V2 model performs the best compared to other two models.

The testing accuracy of three models were calculated using testing data. Table I shows the metrics of three models. The Inception-Resnet-V2 reported the highest training accuracy of 0.88 and testing accuracy of 0.59.

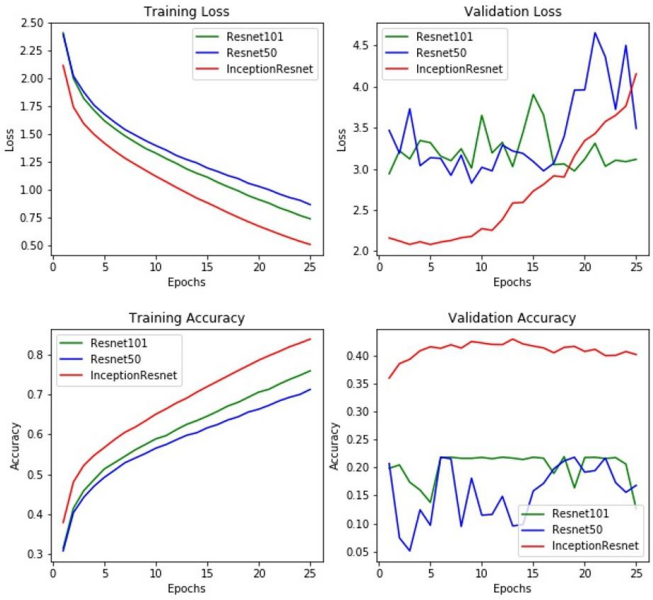

Fig. 3. Loss and accuracy of three models

TABLE I
Model Comparison

models	Accuracy	
	Training	***Testing***
Resnet50	0.71	0.16
Resnet101	0.75	0.09
InceptionResnet v2	**0.88**	**0.59**

B. Gradient-weighted Class Activation Mapping Results

We implemented Gradient-weighted Class Activation Mapping (Grad-CAMs) to inject transparency and explainability into our model. Grad-CAMs produce a localization map which highlights specific areas of the image used by the CNN for classification decisions by utilizing the gradients[11]. It was important for us to interpret the neural network's activation areas in our model for bench-marking our results with their manual counterparts. We also wanted to understand how our model was making decisions which allowed us to fine tune and improve what the model is tasked to do. Furthermore, visualizing activation areas allows domain experts to corroborate the model results with incumbent classification methods. We implemented the methods for a residual network architecture by extracting the activation values from an intermediate convolution layer and using them to generate a heat map identifying the important areas of a given art object image. Figure 5 and 6 show the heat map where our neural networks see. Viewing the images through the Grad-CAMs lens allowed us, and domain experts, to confirm if the model was classifying the images based on real art features that make domain sense or image artifacts. The second is to identify bias when it comes to misclassification in addition to understanding the limitations of our data-set and devise ways to address bias (pre-filtering, up-sampling, down-sampling).

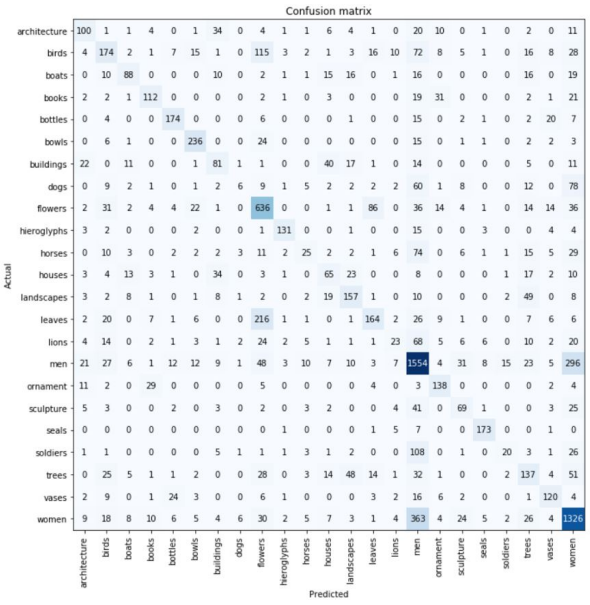

Fig. 4. Confusion Matrix

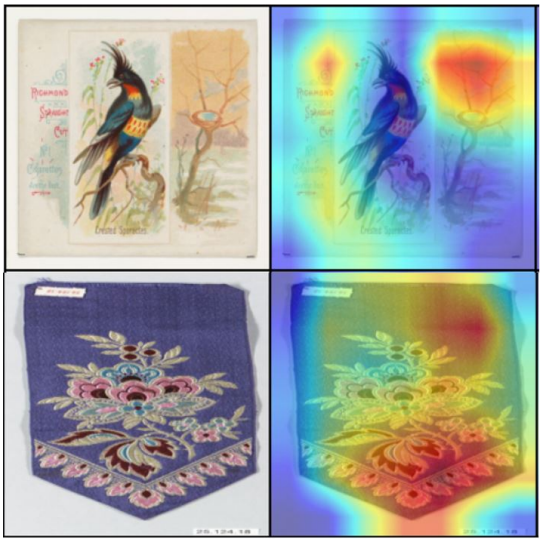

Fig. 5. First row is 'birds' object, and second row is 'flowers' object

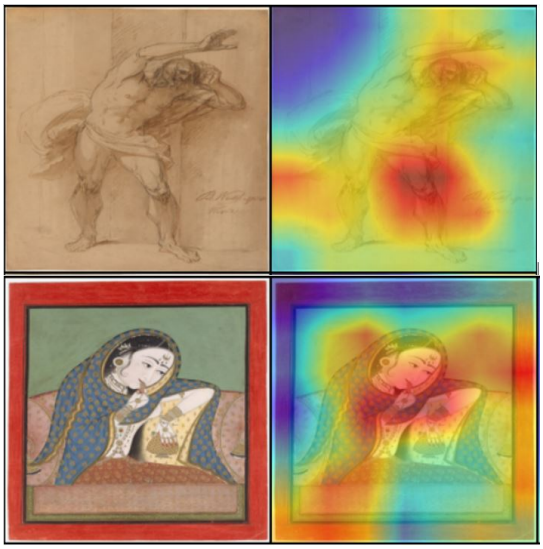

Fig. 6. First row is 'men' object, and second row is 'women' object

Fig. 7. Grad-CAM visualization shows the activation areas which led to the decision produced by our model

Our Resnet50 model accurately labeled 1554 males as males and 1326 females and females. However, the model mislabeled 296 males as females and 363 females as males. We reviewed each and every misclassified object to understand the patterns behind the misclassification and what the visual features (attire, facial features, hair) the model assesses when it makes a decision. In Figure 7, a Native American male with long hair, wearing a traditional regalia including a breechcloth which resembles a dress was incorrectly classified as a female given those specific features.

We noted this to be a common pattern across most of the misclassified gender objects. Males who were misclassified as females were depicted having long hair or were wearing robes, kimonos, or other articles of clothing that resemble feminine attire. For example, Japanese nobility wearing kimonos, Roman men wearing togas, and Mughal emperors wearing a "jama" which are all articles of clothing that resemble long dresses or robes. As evident with the Grad-CAMS, the model was placing importance on parallel lines and on detecting a pattern of long hair which is predominantly a female feature. These mis-classifications bring forth very important issues when working at the intersection of humanities and machine learning especially those that address how we should classify a generalizable model that is suited for all cultural representations and traditions and focuses on facial features when needed. It is important to note that different cultures speak of different human narratives.

C. Multi-label Classification

Results from the model when only the last linear layers were used to train were as follows: after training for one

epoch the model accuracy on the test set was 96.49 percent with a F-score of 0.43 (training loss 0.098, validation loss 0.187100). After training for two epochs the model accuracy on the test set was 96.58 with a F-score of 0.56 (training loss 0.076, validation loss 0.410789). Results from the model all the layers were used to train were as follows: after training for one epoch the model accuracy on the test set was 95.65 percent with a F-score of 0.342 (training loss 0.106479, validation loss 0.819343). After training for two epochs the model accuracy on the test set was 96.42 percent with a F-score of 0.489 (training loss 0.082072, validation loss 0.427356).

VI. Discussion

This paper explores use of CNNs for image classification, compares three different architectures for single object classification and finally uses a multi-classification model. Our results reveal that implementing a multi-label classification model is a more viable and effectual approach than relying on a single-label classification model. We also show that Inception-Resnet-V2 outperforms the other two CNNs which corroborates previous research and documentation.

Interpretation of art can be challenging due to the significant overlap of representation and converging layers of classification. Objects within the collection included a vessel with engravings of flowers and leaves or an embroidered piece of fabric with an abstract bird. It is also important to note that within this space, the medium itself can also be an object within itself. The beauty of art is such that it is not always representing what is apparent to the human eye and has the ability to tell multiple stories from one work. This brought forth questions such as: how do we train the classifier to push it to see what we as humans believe is important when it comes to describing an object? How do we assess the validity and accuracy of the ground truth for human annotators? In terms of broader relevance for art museums and institutions, we note that there are quirks of cataloguing associated to each art museum and establishing a codified process to establish the ground truth labels for a subset of The Met's collection is highly pivotal for ensuring successful integration of AI and Art in the future.

Any attempt to increase user engagement with art will require explorations of different methods of annotation. With the use of human annotators via crowd sourcing platforms (Kaggle, Wikimedia, Amazon Turk, etc.) or via in-house teams, there is always the concern for bias. There has been extensive work done showing how machine learning methods internalize and can often amplify bias. In our work we showed this to be true for bias related to gender through our GRAD-CAM visualizations. Being aware of bias is the first step prior to then addressing it. In fact, at this point given how pervasive gender bias is and how hard a problem this is to solve, industry leaders have chosen to stop using gendered labels like 'woman' or 'man' in photos of people given that the gender of a person cannot be inferred by appearance. The decision was made to align with the Artificial Intelligence Principles at Google, specifically Principle number 2: Avoid creating or reinforcing unfair bias [12].

A natural extension of our research would include pre-filtering the input data to address the nuances in aesthetic traditions. The intuition behind this approach is centered around the notion that aesthetic traditions bring forth commonality between objects which can further inform tag classification. For instance, nobility is depicted very differently in Egyptian art in contrast to Roman art. Given that representations in art abide by cultural norms and traditions of representation, we can look further into segmenting our data-set based on culture. Another alternative approach would include the process of up sampling or down sampling our data-set to add more data representing males with longer hair or wearing dresses to train our model to focus on other descriptive features (i.e. facial hair). Finally, future work would also include use of unsupervised clustering methods / auto-encoders to explore additional themes in the data.

Acknowledgments

We would like to thank our sponsors at The Met (Jennie Choi and Maria Kessler) for providing critical guidance regarding the questions we chose to explore. We also extend our gratitude to our capstone advisor Rafael C. Alvarado for weekly feedback on our progress and to Lane Rasberry, our Wiki consultant for providing context to the open source background of the MET.

Author Information

Logan Yoonhyuk Lee, BS, Graduate Student, School of Data Science, University of Virginia

Sudeepti Surapaneni, BS, Graduate Student, School of Data Science, University of Virginia

Sana Syed, MD MS, Graduate Student, School of Data Science; Assistant Professor of Pediatrics, School of Medicine, University of Virginia

References

[1] Tallon, Loic. "Introducing Open Access at The Met", The Met, 2017, February 7.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. "Imagenet classification with deep convolutional neural networks". In NIPS, 2012.

[3] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In Proceedings of CVPR, pages 770–778, 2016.

[4] Saleh, Babak Elgammal, Ahmed. "Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature" (2015).

[5] Adrian Lecoutre, Benjamin Negrevergne, Florian Yger. "Recognizing Art Style Automatically in Painting with Deep Learning." Proceedings of Machine Learning Research 77:327,342, 2017.

[6] Y. Hong, J. Kim, "Art painting identification using convolutional neural network", International Journal of Applied Engineering Research 12 (4) (2017) 532–539.

[7] Yang, S., Oh, B. M., Merchant, D., Howe, B., and West, J. (2018). "Classifying Digitized Art Type and Time Period". In Proceedings of the 1st Workshop on Data Science for Digital Art History-Tacking Big Data, (2018).

[8] Lecoutre, Adrian et al. "Recognizing Art Style Automatically in Painting with Deep Learning." ACML (2017).

[9] L. Perez and J. Wang. "The effectiveness of data augmentation in image classification using deep learning". arXiv preprint arXiv:1712.04621, 2017.

[10] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. In ICLR Workshop, 2016. 7

[11] Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., 2017, March 21. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”.

[12] Lyons, K. ”Google AI tool will no longer use gendered labels like ‘woman’ or ‘man’ in photos of people”. 2020.