SherlockNet: Exploring 400 Years of Western Book Illustrations With Convolutional Neural Networks

Luda Zhao
Stanford University
ludazhao@cs.stanford.edu

Karen Wang
Stanford University
kywang@cs.stanford.edu

Brian Do
Stanford University
bdo@cs.stanford.edu

1. Abstract

Book illustrations provide valuable insights into the cultural fabric of their time. Traditionally, extraction of factual tags and stylistic trends have required human annotators, making large-scale implementations impractical if not impossible. In this project, we used convolutional neural networks (CNNs) to categorize images and uncover trends in an unbiased manner from Western book illustrations produced between 1500 and 1900. Working with a dataset of over 1 million scanned illustrations from the British Library, we utilized transfer learning techniques to train a classification model for 12 categories (decorations, architecture, animals, etc.) Using a 10K training set, the model achieved 80.6% top-1 accuracy and 96.4% top-3 accuracy on our 12 categories. We then utilized these tags to train CNNs to classify book illustrations based on their dates of publication. We worked with the decorations and maps categories which obtained good results: our model was able to pick the correct publishing era 65.8% and 47.9% of the time, for decorations and maps respectively. We then used neuron visualization techniques to find specific features that became more or less popular over time. Our results demonstrate that CNNs can not only be used as an accurate annotator of illustrations with order-of-magnitude efficiency improvements, but also as a tool to understand large-scale trends and patterns. We hope this work establishes CNNs as a novel tool for annotation and analysis, and encourages further adoption of neural networks in the field of bibliology.

2. Introduction

Book illustrations provide an important window into how social, literary, and artistic constructs have changed over time [2]. As literary tastes and the literate populace evolve, so do the illustrations that accompany the books of the time. Advances in printing technique, changes in styles of art, or simply changing historical context are also reflected in these illustrations, thus making book illustrations a rich and valuable source of historical and cultural information.

Historically, the study of art has proceeded through a bottom-up approach, where trajectories in art are pieced together from intensive study of a small number of hand-picked art pieces [3]. Evaluating artwork one piece at a time is time-consuming, requires great expertise, and relies on humans to recognize patterns that may be subtle or rare. Moreover, the choice of art pieces influences the story that is told. A top-down, computationally guided approach would allow art historians to study the trajectory of art and culture in a more unbiased and efficient manner [9].

Machine learning provides a framework for extracting insights from data on a massive scale. In particular, deep learning methods such as convolutional neural networks (CNNs) are particularly suited for understanding the content of images, due to the highly nonlinear, hierarchical nature of the neurons that make up the network [8]. CNN algorithms have advanced significantly over the past five years, with image classification results on the 1,000 class ImageNet challenge approaching human-level performance [6]. CNNs can also be useful in terms of the image features that it learns. Each layer of a CNN sees a slightly more complex set of features, and these sets are fine-tuned so that they provide maximum separability between the tags that the CNN is classifying over. Methods such as saliency maps [11], dimensionality reduction of codes at each layer, and class optimization provide ways to understand datasets in terms of these intermediate features. In the context of art history, these features can be especially useful because they could theoretically represent styles, patterns, or sets of motifs that are hard for humans to notice. Thus, CNNs could be a useful tool for studying art history via a novel top-down approach.

In 2010, the British Library (through a collaboration with Microsoft Labs) began an initiative to digitize books printed between 1500 and 1900 and make them publicly available. The result was a treasure trove of images ranging from illustrations, decorative motifs, portraits, satirical comics, maps, to geological diagrams that correspond to a period of major expansion in Western book production and popular media. We reasoned that this dataset would provide rich insights into culture and art in the Western world, as well as trends in art styles over time.

In this project we used CNNs to tag the British Library dataset and analyze its historical trends. First, we used a two-step bootstrapping approach to tag these images into one of 12 categories that we deemed to be representative of the dataset. These tags by themselves will provide art historians with a resource for finding works of art that support or extend their current hypotheses. Second, we classified decorations and maps by date and used neuron visualization techniques to find specific features that became more or less popular over time. Together, this work establishes CNNs as a novel framework for gaining insights into the trajectory of book illustrations and culture over time.

3. Approach

3.1. Data Preprocessing

Images and associated metadata (e.g. author, title, publication date) were generously provided to us by the British Library. All images were resized to 299x299 (the image was scaled so that its smaller dimension was 299px, then its larger dimension was cropped to match) and rendered as grayscale.

3.2. Tag Classifier

First, approximately 1,500 images were manually classified into one of 12 categories (animals, architecture, decorations, landscapes, nature, people, miniatures, text, seals, objects, diagrams, and maps). Each category had at least 100 images. These images were fed into a pretrained Inception-V3 convolutional neural network running in TensorFlow on 8 NVIDIA Kepler GK104 GPUs in Amazon EC2 [1, 5, 12]. The last affine softmax layer was retrained for 1,000 minibatch steps (roughly 50 epochs) under various learning rates and using the gradient descent algorithm Adam [7]. Data augmentation was also performed using a series of randomly chosen flips, scales, and crops. This was called the 1.5K model. We randomly partitioned the dataset with a 80%/10%/10% training/validation/testing split.

Using the 1.5K model, 10,000 randomly chosen images were classified into one of the above 12 categories. The tags for these images were then verified manually and corrected as needed. These newly tagged images were then used to train a second model, the 10K model, using similar methods as above. Finally, the 10K model was used to tag all images in the British Library 1 million images dataset.

3.3. Date Classifier + Analysis

Using the generated tags in the last section, we created datasets consisting of single tags. we then trained additional ConvNets on two of these datasets (decorations, and maps, each of size roughly around 100K) to classify the images based on the publishing year of the images. We grouped dates into 7 different buckets as labels(pre-1700s, 1700-1750, 1750-1800, 1800- 1850, 1850-1870, 1870-1890, post-1890s). We used the same Inception-v3 pretrained model and retrained the last affine softmax layer with the new training sets using similar methods as above. Due to the uneven distributions of the number of images per date in the dataset, specific care were taken to ensure the testing set contained even numbers of all classes to ensure non-skewed results.

After training our models on our training set and evaluating results, we turned our attention to the models itself in attempts of inferring trends from the individual neurons in the network. We compiled and analyzed differences in activations of different neurons throughout the network with regards to the different periods. We then used the images with the highest activations of such neurons to generate hypotheses as to the potential interpretations of such differences.

4. Results

4.1. Characterization of Data

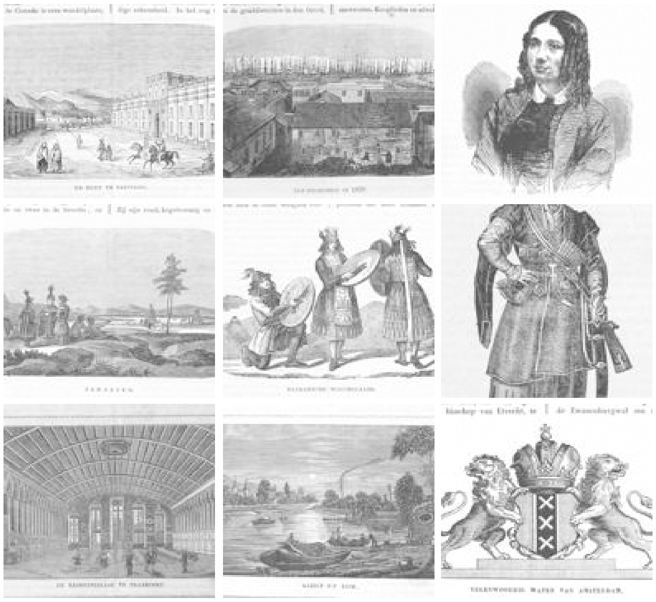

Figure 1: Example pictures from the British Library 1M dataset.

We obtained 1,014,190 images from book scans performed by the British Library (Figure 1). These images are comprised of 414,727 small images (decorative motifs), 216,360 medium images (photos, drawings, etc), and 383,103 larger plate images and come from 30,994 books. The subject matter of these books is varied and extensive. About 70,000 images (7% of the dataset) also have manual

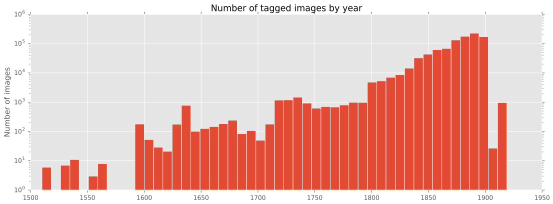

Figure 2: Number of images per year.

tags submitted by Flickr users; however, these tags do not come from a controlled vocabulary.

To further understand the dataset on a global scale, we plotted the number of images (Figure 2). The dataset mostly contains books from 1800 to 1900, with a substantial fraction of those after 1850. While pre-1800 counts are highly variable, the number of images per book generally increased over time, from an average of 15 images per book in 1800 to an average of 50 per book by 1900.

4.2. Image Tagging

As an initial step, we randomly chose 1,408 images from the dataset and manually tagged each of them into one of 12 classes that we deemed to cover most of the dataset (A.E., personal communication). These 1,408 images represented 926 books, had a publication date distribution similar to the overall dataset, and had at least 100 representatives in each class (Figure 3a-b). After manual tagging, we retrained the last layer of Googles Inception-V3 CNN to output softmax scores for our 12 classes. Using stochastic gradient descent (SGD) with a learning rate of 1e-3, we achieved a top-1 validation error rate of 22.0% and a test error rate of 31.7%. Further inspection of class assignments revealed that landscapes, architecture, and nature are often confused with each other, as are miniatures and decorations. In addition, since many images have multiple objects (e.g. people and animals), the choice of tag is often subjective (Figure S1). Thus, we calculated the top-3 error rate (10.4%) and the top-5 error rate (3.9%).

To build a larger training set, we tagged 10,000 randomly chosen images and validated them manually, observing tag inconsistencies similar to what was seen in the manually tagged dataset (Figure 3c). We then retrained the last layer of Inception-V3 using both sets of data. Using the Adam gradient optimizer with a learning rate of 8e-3, as well as data augmentation with 3 transformations per image randomly chosen from a set of predefined flips, crops, and scales, we achieved the following test error rates: 19.7% top-1, 3.6% top-3, and 0.7% top-5 (Figure 4a).

With this model, we classified all images in the dataset and analyzed the global distribution of tags. In total, 970,000 images were tagged, with about 40,000 images dropped due to errors during preprocessing. The dataset is enriched for images of people, with high numbers of landscape, architecture, and decoration images (Figure 4b). Miniatures are rarest, reflecting their presence in select genres of books and only at the start of passages. Most tags follow a similar date distribution as the overall dataset (Figure 4c), with few images prior to 1800 and a monotonically increasing distribution between 1800 and 1900. The majority of pre-1800 images are decorations, with several instances of people, seals, miniatures, and text.

4.3. Date Classification

With a complete dataset of tagged images, we began to explore trends in art styles and content over time. Style trends can often be difficult for humans to distinguish because they can be convolved with content trends, and vice versa. Thus, we reasoned that CNNs might provide a more unbiased approach for study of trends. We decided to focus on Decorations and Maps because their content is relatively homogeneous and because trends would prove useful for study by art historians and cartographers, respectively (Figure 5a). As a first step, we retrained Inception-V3 so that it would be capable of classifying images of each type into specific eras (e.g. pre-1700, 1850-1869, or post-1890). Individual parameter optimization for each (see Approach) yielded a 7-class CNN that achieved an error rate of 34.2% on the Decorations dataset (compared to 86% random chance) and a 4-class CNN that achieved an error rate of 52.1% on the Maps dataset (compared to 75% random chance) (Figure 5b). Confusion matrix analysis showed that for both Decorations and Maps, errors primarily occurred between adjacent eras (Figure 5c-d). While far from perfect, these results indicate that CNNs are able to discover features that at least partially distinguish decorations and maps by the era in which they were produced.

We then set out to characterize the internal structure of our decoration era-specific CNN at four representative layers (Conv0 - the initial convolution layer; Conv4 - the last convolution layer - Mix5 - the fifth Inception layer; and Mix10 - the tenth and last Inception layer, just before the affine softmax layer). For each layer, we ranked the neurons by how much their mean activations differed in the post-1890 set versus the pre-1700 set. A large difference indicates that the neuron is era-sensitive. We found that layers deeper in the network had a smaller proportion of era-sensitive neurons than shallower layers, and that, at least in the deeper layers, there were a roughly equal number of neurons that had increased or decreased activations over time (Figure 6a). This was also true for Maps (not shown).

For the Mix10 layer in the decorations CNN, we additionally plotted the activations of its era-sensitive neurons over time. We found that neurons are not era-specific (Figure 6b) but instead seem to be activated in a contin-

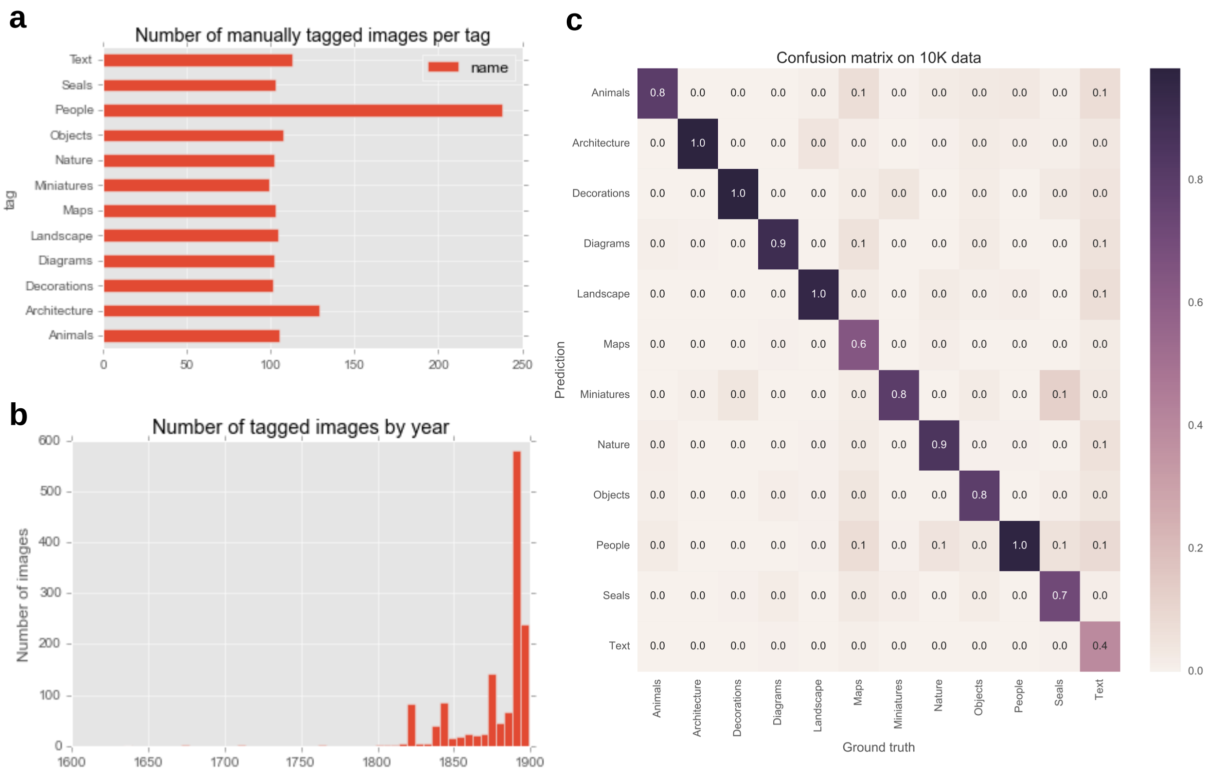

Figure 3: Statistics from manual tagging. a) Number of manually tagged images per tag. b) Date distribution across all manually tagged images. c) Confusion matrix for 10K images after manual validation.

uous manner over eras (Figure 6c). In other words, era-sensitive neurons seem to have activations that smoothly increase or decrease over historical time, as opposed to an all-or-none response with high activations only in one era. We performed the same analysis on shallower layers (Mix5, Conv4, and Conv0) and found a much less pronounced gradient effect. To confirm this, we performed t-SNE on the codes for each image at these layers [13]. Clustering of images into different eras was more pronounced as we moved from Conv0 to Conv4 to Mix5 to Mix10, suggesting that network depth provides discriminatory power between eras (Figure 7). Thus, era-sensitive CNNs have neurons for complex features that emerge or disappear gradually over historical time.

What are these complex features? Because neuron visualization approaches such as deconvolution are still in their infancy in TensorFlow, we decided to interrogate each neurons pattern preferences by surveying images for which this neuron responds most and least strongly. We denote neurons whose activations decrease over time as antique neurons (their activations are highest pre-1700) and neurons whose activations increase over time as modern neurons (their activations are highest post-1890). Images that are most preferred and rejected by antique neurons are displayed in Figure S2 and modern neurons in Figure S3. As an example, the antique neurons 931 and 763 appear to prefer different kinds of small rounded dark objects, which indicates that these motifs are overrepresented in early decorations. On the other hand, the modern neurons 126 and 541 have horizontal borders and complex features within these borders, hinting that these may be common features of more modern decorations. Therefore, visualizing neurons that are more or less activated over time can serve as a generalizable technique for studying art history over time.

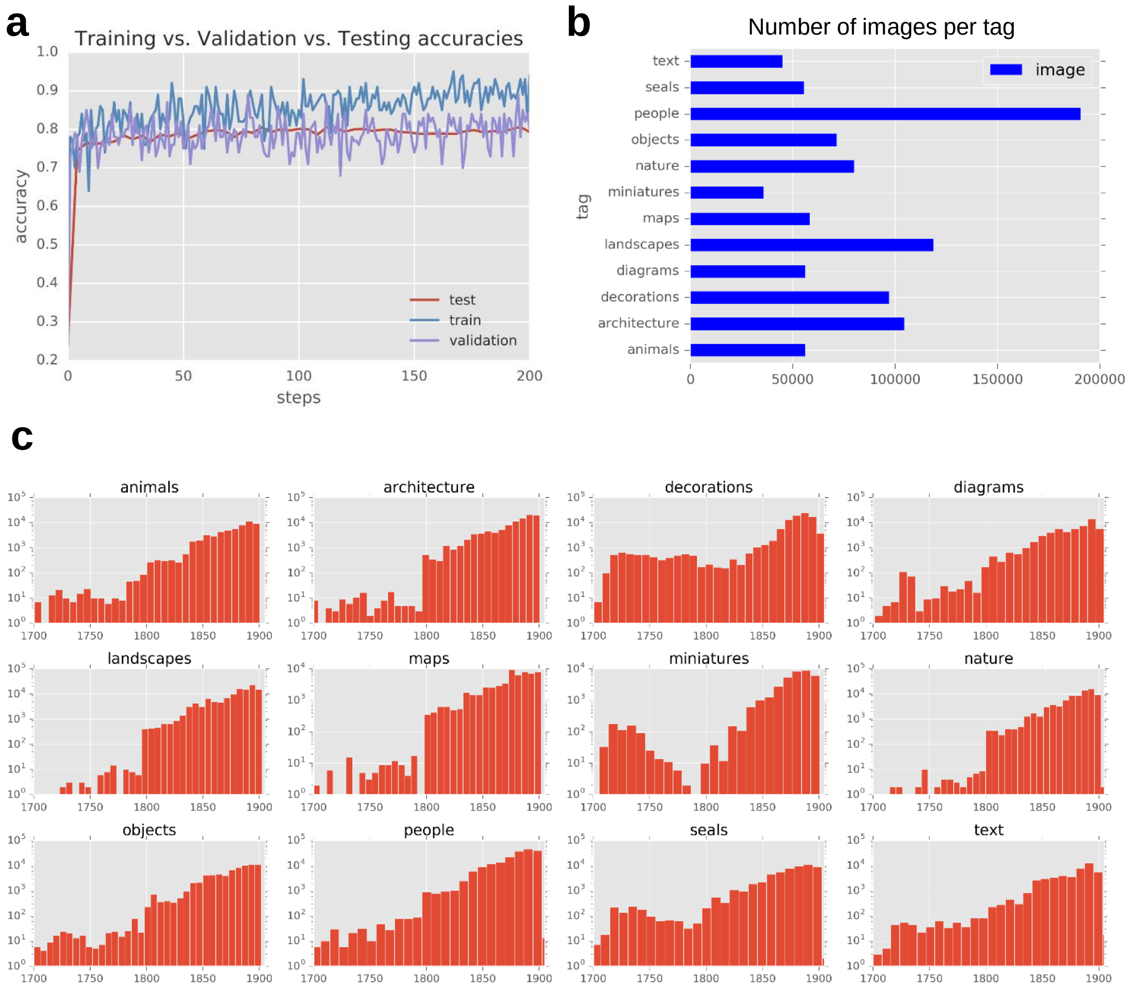

Figure 4: Tagging of all 1 million images. a) Training, validation, and test accuracy for the CNN during training. b) Number of images per tag. c) Date distribution for each tag.

5. Conclusion/Future Work

In this paper we have shown that machine learning and CNNs represent a powerful new technique to studying art history with an unbiased top-down approach. First, we tagged 1 million Western book illustrations from 1500 to 1900 into 12 distinct categories on eight GPUs in less than six hours, achieving 19.7% top-1 and 3.6% top-3 error. Second, we built a CNN that discriminated decorations by era and, by defining neurons as antique or modern based on their activations over time, found specific features of older and newer decorations. These two advances should be of great interest to art historians and serve as a novel approach to hypothesis generation and discovery in the field.

We performed tagging through a two-step bootstrap approach where we first tagged 1,500 images manually and built a small CNN, then used that to tag 10,000 images

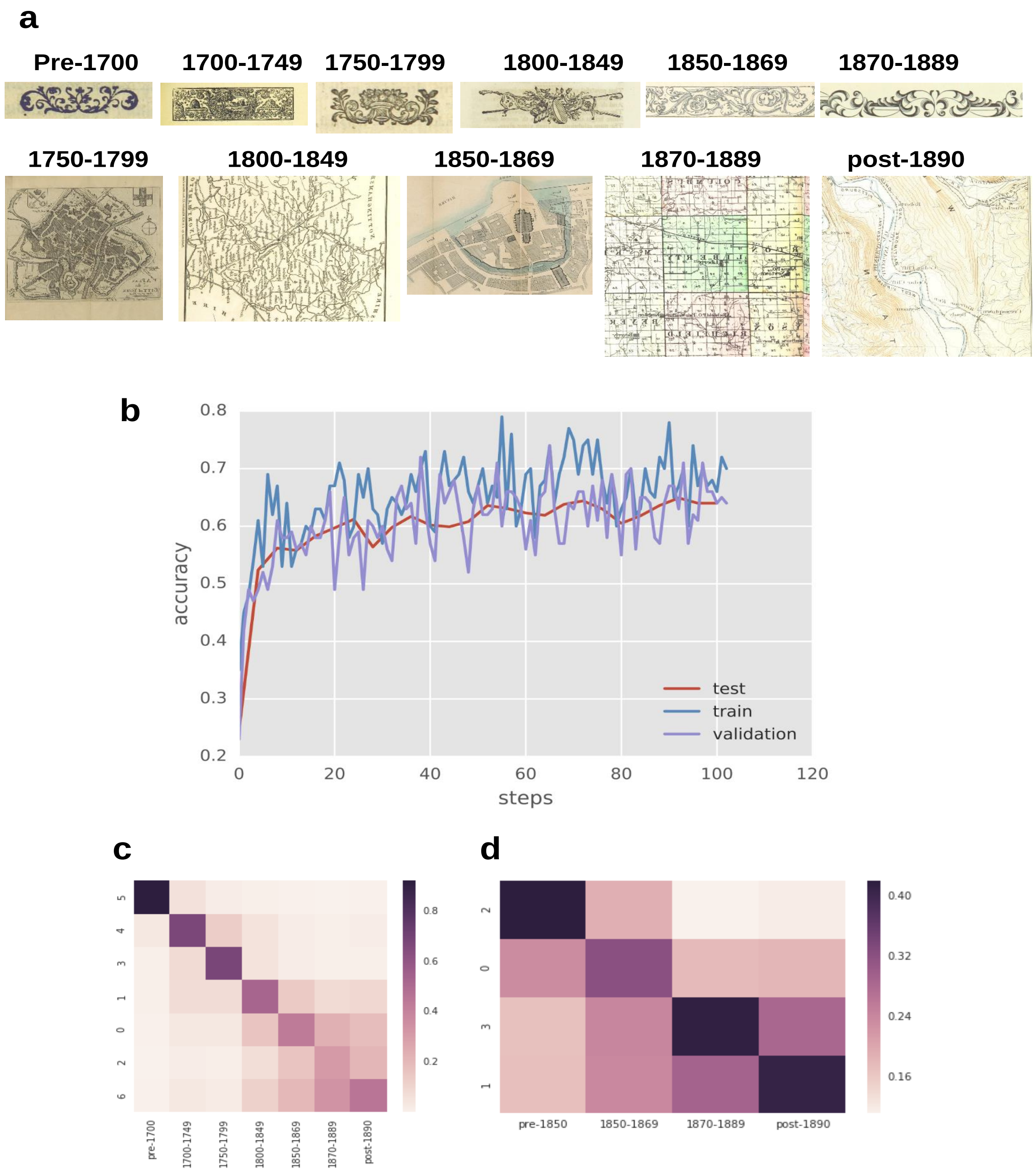

Figure 5: Analysis of decorations and maps over time. a) Examples of decorations (top) and maps over time (bottom). b) Training, validation, and test accuracy for the decoration CNN during training. c) Confusion matrix for decorations. d) Confusion matrix for maps.

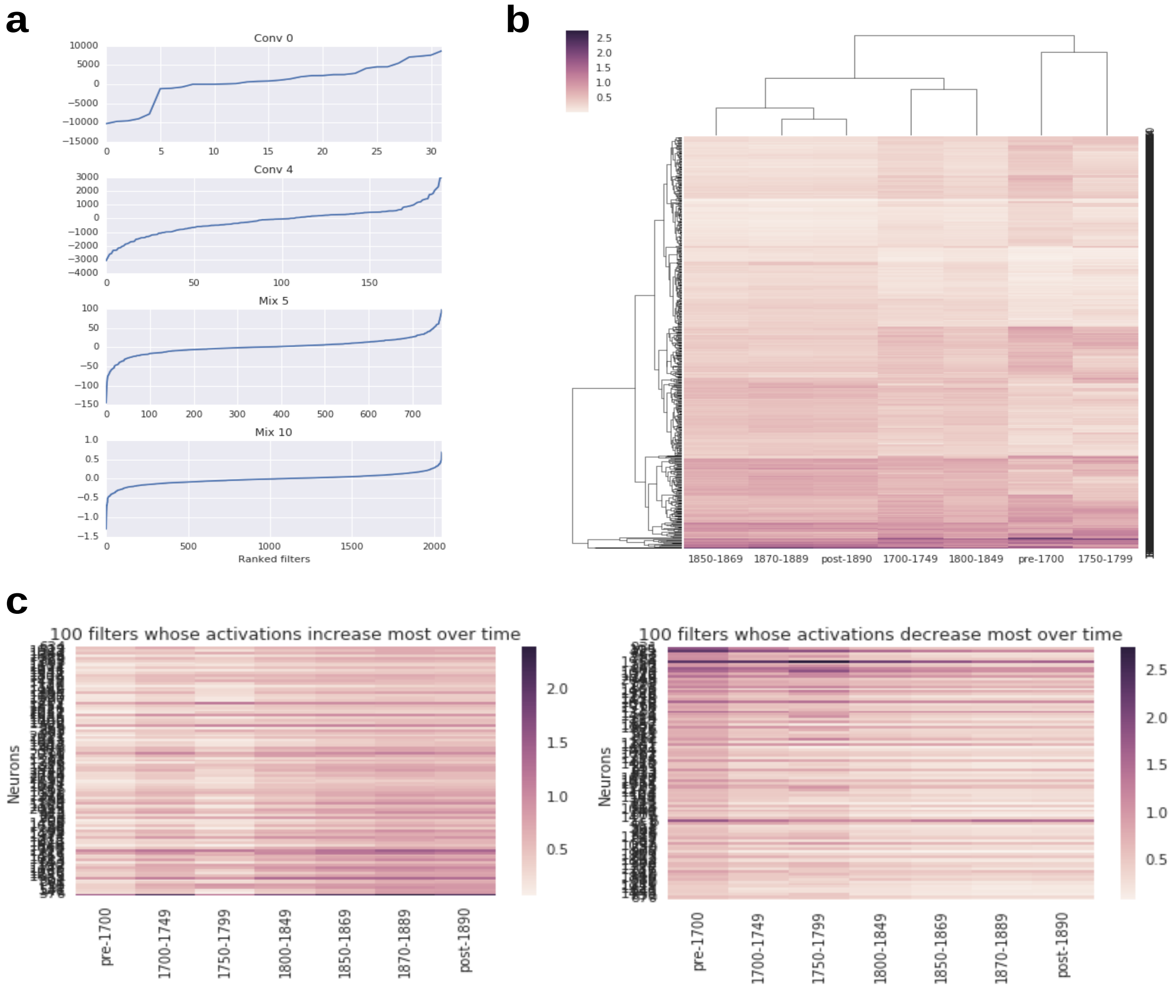

Figure 6: Analysis of neurons in decoration era CNN. a) Distribution of difference between post-1890 and pre-1700 mean activations for all neurons in 4 different layers. Neurons are ranked from left (antique) to right (modern). b) Hierarchical clustering of activations for all neurons in layer Mix10 across all 7 timepoints. Rows represent neurons and columns represent eras. c) Heatmaps of activations for top 100 modern neurons (left) and top 100 antique neurons (right).

which we manually validated and then used to build a larger final CNN. This approach was efficient because tagging 500 images took about 3 hours per person and validating 500 images took less than 15 minutes per person. However, validation of images was more prone to human error than de novo tagging. Additionally, tagging was often ambiguous. Many images, for instance, incorporated both people and animals or diagrams of objects, and it was unclear which tag to assign. For that reason, top-3 accuracy was much more representative of the power of our CNN than top-1 accuracy.

Date classification had much higher error rates than image tagging (see next paragraph for discussion) yet was surprisingly able to classify images into different eras at a rate high above background. This classification performed better for decorations for maps, probably because decorations change quickly whereas cartographic styles are more resistant to rapid change. We leveraged this capability to rank neurons at different layers by their era sensitivity and found that only a small proportion of neurons (especially at deeper

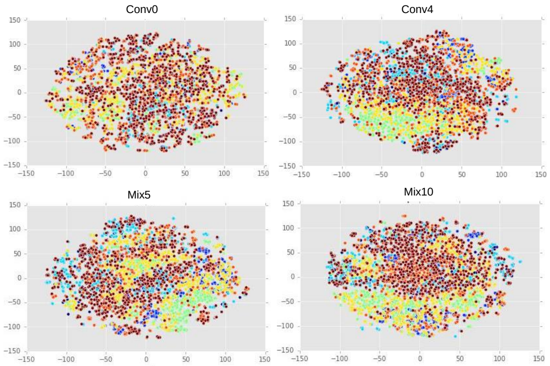

Figure 7: t-SNE visualization of image codes at different layers. Eras are represented as different colors.

layers) was sensitive to era. One interesting finding was that these neurons seemed to be gradually activated or repressed over historical time, rather than being specific to only one era. However, this could have been an effect of averaging different numbers of all-or-none responses at each timepoint rather than smooth decreases in neuronal activations.

We then attempted to visualize neurons that we found were important for distinguishing between eras. One contemporary approach for neuron visualization is to set the gradient for that neuron at 1 with all other gradients set at 0, and continue to backpropagate while updating the input image. However, we encountered difficulties implementing this in a pretrained network in TensorFlow. As an alternative approach, we visually compared images that strongly activate the neuron of interest to images that do not activate the neuron. While this technique succeeded for several neurons, many neurons had uninterpretable preferences (Figures S2-3). We believe that optimization-based neuron visualization would give great insights and are currently working on this.

For both classifiers, we retrained the last layer of Googles Inception-V3. Retraining, or transfer learning, is a powerful method of quickly adapting a pretrained model to specific needs [10, 4]. The assumption is that the higher-order features learned by the model are useful for a wide variety of domains, and only the final affine softmax layer needs to be fine-tuned to the application in question. While retraining worked well in our case for classification into different image classes, we found it performed much less accurately for classification into eras. It is likely that the features that distinguish images by era are likely to be lower-order, so the features that enable class separation (into decorations versus people, for instance) could have been too high-level for era classification.

To conclude, we have used deep learning to tag the British Library 1 million images dataset into 12 categories and discover trends in art styles over historical time. Further work will seek to improve era classification and automatic visualization of artistic and contextual features that change across these eras. Together, these efforts bring art history into the world of machine learning.

6. Acknowledgements

We are grateful to Mahendra Mahey (British Library Labs) for providing the dataset and research contacts, Adrian Edwards (British Library) for historical insights on the data, Elliot Crowley (Oxford) and Mario Klingemann for input on the computational pipeline, and Silviana Ilcus (Stanford) for insights about art history. Funding for computational resources was provided by the CS231N course staff and the Amazon Web Services Educate program.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] M. Bland. *A Guide to Early Printed Books and Manuscripts*. Wiley, 2010.

[3] K. Bowen and D. Imhof. *Christopher Plantin and Engraved Book Illustrations in Sixteenth-Century Europe*. Cambridge University Press, 2008.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.

[5] Google. Inception-v3. `https://github.com/google/inception`, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[8] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. Insight.

[9] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, , J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[10] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.

[11] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[13] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. 2008.