transcript Verlag
Bielefeld University Press

Chapter Title: Artificial Intelligence and Discovering the Digitized Photoarchive
Chapter Author(s): X.Y. Han, Vardan Papyan, Ellen Prokop, David L. Donoho and C. Richard Johnson <suffix>Jr.</suffix>

Book Title: Archives, Access and Artificial Intelligence
Book Subtitle: Working with Born-Digital and Digitized Archival Collections
Book Editor(s): Lise Jaillant
Published by: transcript Verlag, Bielefeld University Press. (2022)
Stable URL: https://www.jstor.org/stable/jj.11425482.4

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

This book is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/. Funding is provided by Universität Bern, Loughborough University.

transcript Verlag, Bielefeld University Press are collaborating with JSTOR to digitize, preserve and extend access to *Archives, Access and Artificial Intelligence*

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Chapter 1: Artificial Intelligence and Discovering the Digitized Photoarchive

X.Y. Han, Cornell University | Vardan Papyan, University of Toronto | Ellen Prokop, National Gallery of Art, Washington, DC | David L. Donoho, Stanford University | C. Richard Johnson, Jr., Cornell University.

Abstract[1]

This chapter introduces the technical aspects of a useful model for effective interaction between the fields of computer science and cultural heritage preservation. Machine learning researchers at Cornell University, Stanford University, and the University of Toronto are collaborating with staff members of the Frick Art Reference Library (FARL), New York, to explore how computer vision might enhance the accessibility and discoverability of the Library's digital resources. Focusing on FARL's Photoarchive—a research collection of 1.2 million reproductions of works of art in the Western tradition—we are seeking to leverage recent tools in artificial intelligence (AI) to automatically annotate images in the collection with the headings used in the Photoarchive's local, iconography-based classification system. This is being achieved by engineering the syntax of the local classification system into the training and predictive process of deep convolutional neural networks, the cornerstone of modern AI advancements. Thus, the machine leaning researchers are adapting state-of-the-art AI techniques by drawing on and incorporating the expertise of art historians. We demonstrate promising performance metrics and offer informative scientific insights that have the potential to create a valuable tool for metadata creation and image retrieval, an end-product that will address the real-world challenge faced by FARL staff who must manage a growing backlog of images that are digitized but not yet classified.

Section 1. Introduction and Background

Computer vision and computational art history are naturally synergistic fields. With the internet's growing presence in modern life, art museums and cultural heritage institutions are digitizing their collections to connect to a wider and more

1 X.Y. Han, Ellen Prokop, and Vardan Papyan contributed equally to this chapter and are listed alphabetically.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

inclusive audience through their websites and social media platforms and thus are releasing terabytes of high-quality, annotated digital images online. Meanwhile, state-of-the-art deep neural networks have achieved near human-level performance in the identification of the subject matter and formal qualities of digitized images, a performance predicated on the availability of large and fully labeled training datasets such as those produced by museum and art library staff. Therefore, while the union of computer vision and art history may at first appear surprising, it offers great benefits to both disciplines. In this chapter, we document a collaboration between computer vision researchers and art librarians to harness recent advancements in artificial intelligence (AI) and machine learning (ML)[2] to develop an algorithm that has the potential to become a valuable tool for metadata creation and image retrieval. While we do develop new, specialized technical tools for this task (see Section 2.4), we emphasize the notability of this project is instead defined by the successful integration of the expertise of both the art history and computer science communities—through a collaborative back-and-forth that has now been ongoing for more than three years (since 2017)—into a pipeline that possesses mutually-acknowledged practicality.

Section 1.1 The Frick Art Reference Library's Photoarchive

The Frick Art Reference Library (FARL) was established in 1920 by the philanthropist Helen Clay Frick (1888–1984) (Fig 1.1) as a memorial to her father, the industrialist, financier, and art collector Henry Clay Frick (1849–1919).[3]

2 The distinction between AI and ML is vague—even to computer scientists. AI often refers to more modern, end-to-end algorithms in which the input is the raw data and the output is a recommended decision. In contrast, ML often refers to a broader class of computational algorithms that may first require human pre-processing of the raw data (often based on mathematical principles) before applying a main algorithm that may either output the final decision or just a subtask of the decision. Yet this distinction merely describes the contexts in which such terms are applied; in practice, the terms are often used interchangeably.

3 Martha Frick Symington Sanger, *Henry Clay Frick: An Intimate Portrait*, New York 1998, 499.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Fig 1.1: Portrait of Helen Clay Frick in her office at the Frick Art Reference Library, 1939, photographer unknown. Frick Family Photographs. Courtesy of The Frick Collection/Frick Art Reference Library Archives.

The founding collection of this research institution is the Photoarchive, a study collection of more than 1.2 million reproductions of works of art in the Western tradition from the fourth through the twentieth centuries, a resource modeled on the famous Library of Reproductions assembled by Sir Robert Witt (1872–1952) and his wife and housed at their home at 32 Portman Square in London.[4] The Witts' archive served as a "central and comprehensive storehouse [of reproductions] for easy and rapid reference and research" and was open by appointment to "scholars, critics, writers, collectors, [and] dealers."[5] When Helen Clay Frick visited the Witts' archive in the summer of 1920, she immediately recognized its value for art historical scholarship and determined to assemble a similar research collection in North America.[6] With the foundation of her archive, Helen Clay Frick sought to advance the study of art history in the United States, the development of which had been

4 Katharine McCook Knox, *The Story of the Frick Art Reference Library: The Early Years*, New York 1979, 6–7. For information on Sir Robert Witt, see: Sir Robert Witt, Dictionary of Art Historians, URL: https://arthistorians.info/wittr [last accessed: April 2, 2021].

5 Knox, *The Story of the Frick Art Reference Library*, 7.

6 Knox, *The Story of the Frick Art Reference Library*, 13.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

impeded by several factors, the most significant of which was that high-quality photographs of works of art were often prohibitively expensive for many students and scholars.[7] During the nineteenth and early twentieth centuries, North American art historians often had to travel to complete their research, which was an option that only a few could afford. Photoarchives such as those founded by Helen Clay Frick, which made possible the consultation of hundreds of high-quality reproductions with related documentation at one time, helped not only to promote the discipline in the United States but also to motivate critical developments in the field, shifting the focus of study from artist biographies to comparative analysis.[8]

To expand the accessibility of this research collection, FARL staff began digitizing the Photoarchive in the late 1990s, a project that will continue through at least 2025. By the fall of 2020, digital images and documentation for more than 317,500 works of art have been made freely available on the institution's online digital archive, The Frick Digital Collections.[9] In December 2022, approximately 230,000 additional images are scheduled to be uploaded. Therefore, in the next two years, The Frick Digital Collections will potentially host images and metadata representing more than 547,500 works of art.

Unfortunately, cataloguing and metadata creation are not keeping pace with digitization and the backlog of images that have been digitized but are not yet catalogued is growing rapidly. FARL's photoarchivists decided that the accessibility and discoverability of the Photoarchive's holdings is the institution's priority and thus determined that all digital images will be released online with only minimal documentation. This includes the artist, title of the work of art, date of execution, and the institution's local, iconography-based classification system, which not only provides a fixed filing location for each reproduction in the physical archive but also increases the online discoverability of specific subjects and themes. Staff will enhance the online catalogue once the entire research collection has been digitized. Yet even applying minimal information for each digitized image is a time-consuming process.

FARL's photoarchivists investigated crowdsourcing as one means to augment the rate of metadata creation. Preliminary experiments with crowdsourcing, however, were unsatisfactory. For many volunteers, applying the Library's local classification system (described in Section 1.2) proved too restrictive and they either

7 Knox, *The Story of the Frick Art Reference Library*, xi.

8 For additional information about FARL's Photoarchive and its impact on the development of art history in the United States, see: Ellen Prokop, Digital Art History for the Masses? The Role of the Public Digital Art History Lab, in: *Život umjetnosti: Journal for Modern and Contemporary Art and Architecture* 105 (2/2019), 196–213.

9 The Frick Digital Collections are available at: https://digitalcollections.frick.org/ [last accessed: April 2, 2021].

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

abandoned the project or resorted to labeling the images with their own terms. These tags as developed and applied by volunteers certainly increase the discoverability of images in the digital realm but they do not provide standardized search results, which is of paramount concern for scholars who require all known examples of a certain subject or compositional element when conducting their research. Another solution was necessary. Therefore, Library staff launched a pilot project in collaboration with a team of computer vision researchers from Cornell University, Stanford University, and the University of Toronto[10] to bring both efficiency and standardization to the cataloguing process using new advancements in AI.[11] The computer vision researchers were intrigued by both the Photoarchive's holdings, which offered a unique dataset, and the classification system used to organize this research collection (described below), which afforded an exceptional intellectual challenge.

Section 1.2 Organization of the Photoarchive

As noted above, FARL's Photoarchive is modeled on Sir Robert Witt's photo study collection, which was deeded in 1944 to the University of London and later incorporated into the research libraries of The Courtauld Institute of Art.[12] The Witt Library as it is currently known features hundreds of thousands of photographs and published reproductions of works of art mounted on archival-quality paper. The artist, title, and date of the work of art are noted on the front of the sheets; on occasion, additional information regarding the attribution of the work or its provenance—that is, its record of ownership—is included on the reverse. These sheets or "photo study mounts" are stored in large boxes organized alphabetically by artist

10 When the collaboration began in 2017, X.Y. Han, Vardan Papyan, and David L. Donoho were at Stanford University while C. Richard Johnson, Jr. was at Cornell University and serving as FARL's Senior Research Advisor. Since that year, Han has moved to Cornell and Papyan to the University of Toronto—leading to the wide geographic spread of this project.

11 During the term of the collaboration, Han was supported in part by National Science Foundation Grants DMS-1407813, DMS-1418362, and DMS-1811614, and donations to Cornell University from private donors; Papyan was supported by National Science Foundation Grants DMS-1407813, DMS-1418362, and DMS-1811614, the Koret Foundation, and other private donors; Donoho was supported by National Science Foundation Grants DMS-1407813, DMS-1418362, and DMS-1811614, and by donations to Stanford University from Anne T. and Robert M. Bass; Johnson was supported in part by the National Science Foundation Grant CCF-1822007 and donations to Cornell University from Geoffrey and Susan Hedrick and other private donors. FARL donated staff time to the initiative; any travel-related expenses were supported in part by funds from private donors.

12 Witt Library, The Courtauld, URL: https://courtauld.ac.uk/study/resources/image-libraries/witt-library [last accessed: April 2, 2021].

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

and grouped in national schools; within the boxes, the artist's oeuvre is subdivided by subject to aid discoverability.[13]

When assembling FARL's Photoarchive, Library staff purchased black-and-white photographs from agents and dealers or cut reproductions from sales catalogues and mounted these images on nine-by-twelve-inch sheets of archival-quality grey cardboard. On the front of the sheets, they recorded the artist or national school, the title or subject, the current location, medium, and dimensions of the work of art. On the reverse, they sought to provide more complete documentation than the Witts and included information regarding the object's date of execution, attribution history, exhibition history, conservation history, provenance, and physical characteristics. Additional data such as the source of the mounted photograph or reproduction, documentation of other sources of reproductions, and a bibliography might also be noted. Like the photo study mounts in the Witts' archive, the FARL mounts were grouped by national school, filed alphabetically by artist, and then subdivided by subject. (If the artist was unknown, the work was filed under the national school and subject only.) Helen Clay Frick and her staff, however, applied a numerical classification system to the subject categories, one that incorporated the artist's national school. This improvement to the Witt's system resulted in a stable filing position for each mount and allowed for the discoverability of specific subjects and themes.[14]

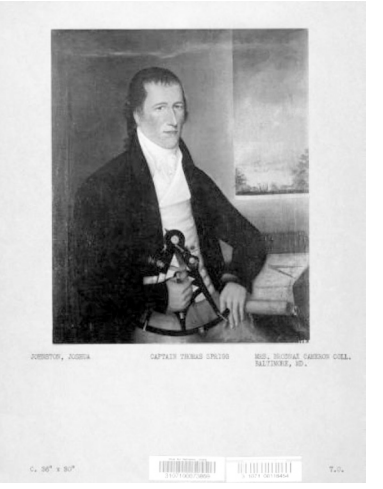
Thus, the half-length portrait of Captain Thomas Sprigg (ca. 1765–1810) by the American artist Joshua Johnson (ca. 1763–1824) is not filed under the artist's name in a folder labeled "portraits" as it would be in the Witt Library but catalogued under the artist's name with the classification heading "121-6" (Fig 1.2 and 1.3).

13 Knox, *The Story of the Frick Art Reference Library*, 17–18.

14 Knox, *The Story of the Frick Art Reference Library*, 17.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Fig 1.2: Photo study mount of Joshua Johnson's Captain Thomas Sprigg (ca. 1805–1810), obverse. Courtesy of the Frick Art Reference Library Photoarchive.



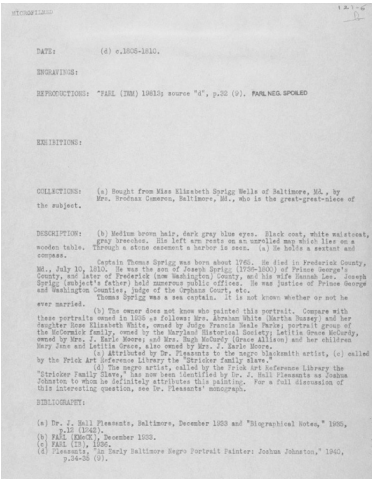
The breakdown of this heading is as follows: the first "1" in the series designates the American School; the following "21" denotes a portrait of a man; and the "6" included after the dash indicates a half-length subject not wearing a hat facing right (as opposed to one facing left, which would be indicated with a "7," or one mounted on a horse, which would be indicated with a "3"). A group portrait of a man and two boys such as Johnson's portrait of John Jacob Anderson and his sons John and Edward presently in the collection of the Brooklyn Museum[15] is classified as "124-7" ("American School: Portrait Groups: Men and Children"); however, Johnson's charming portrait of the three sons of the Westwood family preserved in the collection of the National Gallery of Art in Washington, DC[16] is classified as "124-4" ("American School: Portrait Groups: Children"). The engineering process to adapt modern AI techniques to automatically classify images labeled with the Library's in-house classification system is described in the next section.

15 Joshua Johnson, John Jacob Anderson and Sons, John and Edward, The Brooklyn Museum, URL: https://www.brooklynmuseum.org/opencollection/objects/2168 [last accessed: April 2, 2021].

16 Joshua Johnson, The Westwood Children, National Gallery of Art, URL: https://www.nga.gov/collection/art-object-page.45955.html [last accessed: April 2, 2021].

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Fig 1.3: Photo study mount of Joshua Johnson's Captain Thomas Sprigg (ca. 1805–1810), reverse, with classification number at upper right. Courtesy of the Frick Art Reference Library Photoarchive.



DATE: (d) c.1805-1810.

ENGRAVINGS:

REPRODUCTIONS: "FARL (TRM) 19613; source "d", p.32 (b). FARL NEG. SPOILED

EXHIBITIONS:

COLLECTIONS: (a) Bought from Miss Elizabeth Sprigg Wells of Baltimore, Md., by Mrs. Brodnax Cameron, Baltimore, Md., who is the great-great-niece of the subject.

DESCRIPTION: (b) Medium brown hair, dark gray blue eyes. Black coat, white waistcoat, gray breeches. His left arm rests on an unrolled map which lies on a wooden table. Through a stone casement a harbor is seen. (a) He holds a sextant and compass.

Captain Thomas Sprigg was born about 1745. He died in Frederick County, Md., July 10, 1810. He was the son of Joseph Sprigg (1720-1800) of Prince George's County, and later of Frederick (now Washington) County, and his wife Hannah Lee. Joseph Sprigg (subject's father) held numerous public offices. He was justice of Prince George and Washington Counties, judge of the Orphans Court, etc.

Thomas Sprigg was a sea captain. It is not known whether or not he ever married.

(b) The owner does not know who painted this portrait. Compare with these portraits owned in 1938 as follows: Mrs. Abraham White (Martha Busey) and her daughter Rose Elizabeth White, owned by Judge Francis Neale Parke; portrait group of the McCormick family, owned by the Maryland Historical Society; Letitia Grace McCurdy, owned by Mrs. J. Earle Moore; and Mrs. Hugh McCurdy (Grace Allison) and her children Mary Jane and Letitia Grace, also owned by Mrs. J. Earle Moore.

(c) Attributed by Dr. Pleasants to the negro blacksmith artist, (c) called by the Frick Art Reference Library the "Stricker family slave."

(d) The negro artist, called by the Frick Art Reference Library the "Stricker Family Slave," has now been identified by Dr. J. Hall Pleasants as Joshua Johnston to whom he definitely attributes this painting. For a full discussion of this interesting question, see Dr. Pleasants' monograph.

BIBLIOGRAPHY:

(a) Dr. J. Hall Pleasants, Baltimore, December 1933 and "Biographical Notes," 1938, p.18 (TS42).
(b) FARL (KNeCE), December 1933.
(c) FARL (IT), 1936.
(d) Pleasants, "An Early Baltimore Negro Portrait Painter: Joshua Johnston," 1940, p.14-28 (0).

Fig 1.4: Example of hierarchical structure of the FARL classification system for the "Mythological" heading. The left shows the original heading in the FARL system. The right shows the same heading represented as a tree diagram.

15 Mythological 15-1 Mythological: Single figures 15-2 Mythological: Single figures with amoretti 15-3 Mythological: Groups: Four figures and fewer 15-4 Mythological: Groups: More than four figures 15-5 Mythological: Amoretti	Mythological: +---> Single figures: | +---> With amoretti | | \---> ~ | \---> ~ +---> Groups: | +---> Four figures and fewer | | \---> ~ | +---> More than four figures | | \---> ~ | \---> ~ +---> Amoretti | \---> ~ \---> ~

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Section 1.3 AI for Image Classification

This pilot project builds upon the deep convolutional neural networks algorithm that has made rapid leaps in performance in the past decade: the study of this powerful algorithm is often called "deep learning."[17] By passing images through numerous computational layers, deep networks extract relevant features from images in order to predict the category to which the image belongs. Their behavior is controlled by a set of parameters—often numbering in the millions. In order to achieve the best performance, empirical algorithms must adjust these parameters by "training" them on a large collection (usually with at least tens of thousands of images) of pre-labeled images. More details are found in Section 2.

When large training sets exist, deep nets have become valuable scientific tools for performing complicated predictive tasks. For example, they have been employed by doctors to identify cancer cells[18] and by physicists to discover new subatomic particles.[19] The FARL's Photoarchive is perfectly positioned to benefit from this technology. When the collaboration was launched, FARL staff had digitized and catalogued 57,803 reproductions of American paintings, the majority of which were portraits and represented 642 unique classification headings and subheadings.[20] These tens of thousands of labeled-and-digitized images represent an ideal training dataset, and the trained network can be deployed to expedite the annotation of the hundreds of thousands of digitized images in the Photoarchive that have yet to be labeled.

Moreover, the Photoarchive presents an interesting engineering challenge since its dataset differs from conventional deep learning classification tasks in two ways. First, each classification heading consists of a series of constituent descriptors. Second, the Photoarchive classification headings follow a hierarchical structure (see Fig 1.4). For example, when photoarchivists applied the classification heading "121-6" or "American School: Portraits: Men: With hands (without hats): Head to right" to the half-length portrait of Captain Sprigg introduced above, they considered the descriptor "Head to right" only after the components "Portraits" and "Men" had already applied. (See Section 1.2.)

To tackle this problem, FARL staff and the artificial intelligence researchers at Cornell University, Stanford University, and the University of Toronto collabo-

17 Ian Goodfellow/Yoshua Bengio/Aaron Courville, *Deep Learning*, Cambridge 2016.

18 Dayong Wang et al., Deep Learning for Identifying Metastatic Breast Cancer, *arXiv preprint*, URL: arxiv:1606.05718 [last accessed: April 2, 2021].

19 Pierre Baldi/Peter Sadowski/Daniel Whiteson, Searching for Exotic Particles in High-Energy Physics with Deep Learning, in: *Nature Communications* 5 (1/2014), 1–9.

20 These reproductions were among the first archival resources to be digitized by Photoarchive staff because the Library owns the copyright to these photographs.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

rated to develop a deep learning framework specialized to the Photoarchive's local classification system. The photoarchivists developed a decision tree capturing the classification system's syntax, and the deep learning researchers incorporated this syntax into both the objective with which the network parameters are optimized as well as the algorithm with which the network's decision is made.

In the latest version, we modified the popular ResNet152 network[21]—with 152 layers and 6 million parameters—and trained it on 46,242 classified reproductions of American paintings from the collection provided by the FARL's Photoarchive described above (see Section 3.1). More empirical and engineering details are provided in Section 3.

After training, the network was fed images from the unlabeled portion of the Library's Photoarchive and the network predicted a classification heading for each one. (For more details about the dataset, see Section 3.) These images were then annotated with the predictions and shown to the Library's photoarchivists through an application developed using the crowd-sourcing platform, Zooniverse.[22]

Section 1.4 Human Validation with Zooniverse

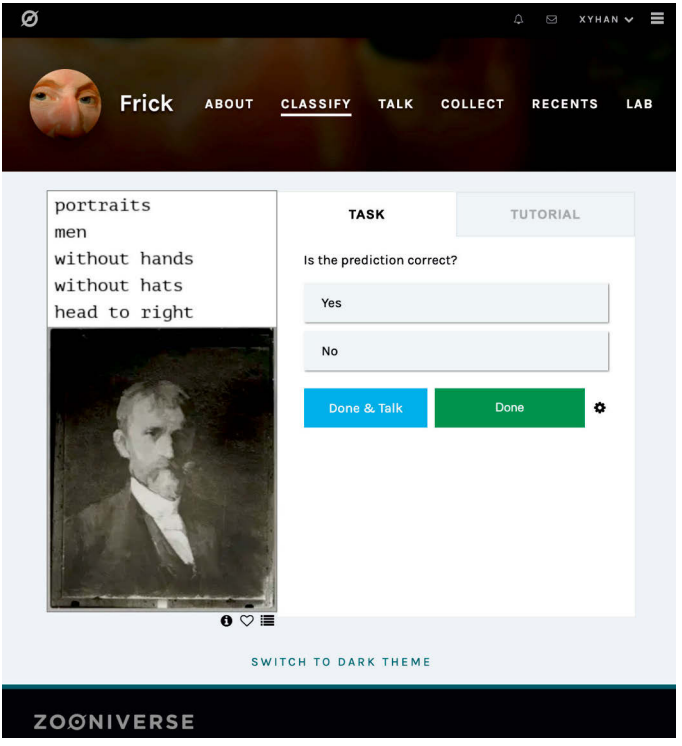
Zooniverse is a popular "citizen science" website where research teams post raw data in need of human annotation and processing. Through the Zooniverse desktop and mobile app, volunteers from the general public or specially-selected groups can then assist with metadata creation. When developing this app, we sought to produce an intuitive interface that would allow archivists to review hundreds of images quickly and easily; thus, the program mimics the notorious dating app Tinder. Library staff downloaded the app on their computer desktops or smartphones (Fig 1.5 and 1.6) and reviewed the algorithm's predictions. If the classification heading applied to an image was correct, it was considered a "match" and the staff member swiped right. If it was incorrect, staff swiped left and the image was sent to a folder for review.

21 Kaiming He et al., Deep Residual Learning for Image Recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 770–78.

22 Zooniverse, URL: https://www.zooniverse.org [last accessed: April 2, 2021].

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Fig 1.5: Screenshot of the application developed using the crowd-sourcing platform Zooniverse to vet the algorithm's predictions as it appears on a computer desktop. Courtesy of the authors.



This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Fig 1.6: Screenshot of the application developed using the crowd-sourcing platform Zooniverse to vet the algorithm's predictions as it appears on a smartphone. Courtesy of the authors.

In testing the latest version of the network, photoarchivists vetted 8,661 images in the year August 2019–August 2020 and agreed with the network on the classification heading for 5,829 (67%) of the images. Yet even the incorrect predictions were, in general, "almost correct" with only one descriptor incorrect or missing. For example, the algorithm applied the classified heading "124-2: American School: Portrait Groups: Men" when it should have applied the heading "124-7: American School: Portrait Groups: Men and Children." In the future, we predict that the network can "learn" to identify these cases when given additional training examples of the headings on which it erred. In Fig 1.9, we show a small selection of outputs from

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

the network that demonstrate the types of one-term omission and extra inclusion mistakes that the network tends to make.

Fig 1.9: An example of typical outputs from the Zooniverse app on images where the network applies the incorrect label. We see that errors tend to be one-term omissions or extra inclusions. Moreover, notice that, to ensure the quality of the annotations, each image is vetted by multiple FARL personnel (each with a unique "user_name"). Their decision is indicated by "annotation" and their correction is given in "comment_body." The "expert_status" column indicates whether the user is a FARL staff member ("expert") or intern ("non-expert"). Attesting to the reliability of the process, most annotators tend to be in agreement on the required correction. Courtesy of the authors.

image_name	network_prediction	annotation	comment_body	user_name	expert_status
31071001117686_001.jpg	portraits: women: with hands: head to right:	No	Portraits: Women: Without hands: Head to right.	JohnMcQ	expert
31071001117686_001.jpg	portraits: women: with hands: head to right:	No	Portraits: Women: Without hands: Head to right.	sarahbigler	expert
31071001120651_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	JohnMcQ	expert
31071001120651_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	kerrip	expert
31071001120651_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	sarahbigler	expert
31071001120651_001.jpg	portraits: men: without hands: without hats: head to right	No	Portraits: Men: With hands (without hats): Head to right.	Ldla657	non-expert
31071001127292_001.jpg	miniatures	No	Miniatures: Portraits: Men.	JohnMcQ	expert
31071001127292_001.jpg	miniatures	No	Miniatures: Portraits: Men.	sarahbigler	expert
31071001128588_001.jpg	portraits: men: with hands: without hats: head to right	No	Portraits: Men: With hands: With hats: Head to right.	levadas	expert
31071001128588_001.jpg	portraits: men: with hands: without hats: head to right	No	Portraits: Men: With hands: With hats: Head to right.	sarahbigler	expert
31071001131452_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats.	levadas	expert
31071001131452_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats.	kerrip	expert
31071001131452_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats.	sarahbigler	expert
31071001131452_001.jpg	portraits: women: without hands: head to left:	No	Portraits: Women: Without hands: With hats	__genie	non-expert

As the digitization of the Photoarchive proceeds, additional image sets will become available for further testing, the results of which will also be vetted by Library staff using the app described above. The team anticipates that the network will be ready to be deployed within two years; thus, automatic image classifiers will become available for cataloguing purposes by 2022.

This project contributes to the growing field of works applying computer vision techniques to art classification and cultural heritage preservation. We discuss some of these related works in Section 4. Yet, this work is set apart by the back-and-forth collaborative process between FARL staff and the researchers at Cornell University, Stanford University, and the University of Toronto, which is creating a high-performing image classifier that follows the syntax of the FARL in-house classification system. In the following sections, we describe in detail the design of the image classifier as well as scientific measurements of its performance.

Section 2. Image Classification with Deep Nets

In image classification, we are given a training dataset of n example pairs, $\mathcal{D} \equiv \left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, where x_i denotes the i-th image and y_i, called the *target*, is typically a term from a controlled vocabulary. For example, for the FARL classification system, we had total of 57,803 images and the descriptions were chosen from the

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

FARL vocabulary of 378 terms, such as "portrait," "landscape," etc. Such a dataset documents each of the descriptions that human experts would provide were they to describe the corresponding images. We seek to learn from this data a computational procedure that can successfully reproduce the descriptions from human experts on this dataset and hopefully on other similar datasets. Because the image descriptions come from a controlled vocabulary, experts, in describing an image, in effect classify it into one of several categories. So, a procedure which reproduces expert judgements can be called an image classifier.

Today's standard for image classification—a modern deep network—involves a *feature extraction* pipeline consisting of many *layers* jointly trained to distill relevant features of the input image. The output of this pipeline is a vector of *scores*, which determine the network's final description. Conceptually, each such pipeline computes, for some input image x, a function $f(x)$ giving those scores. Neural networks map the scores into descriptions by applying some fixed decision function $d(\cdot)$; each score vector $f(x)$ will induce the specific description

$$\hat{y}(x) \equiv d(f(x)),$$

called the *network prediction*.

In practice, the many layers of a deep network contain adjustable parameters, and we must *train* or *learn* these parameters to get a useful performance. Let θ denote the collection of adjustable parameters; f_θ will denote the scoring function f, when those parameters take the specific value θ.

Machine learning algorithms attempts to minimize the cross-training-set average of a *loss function*, $L(\cdot,\cdot)$, on $\mathcal{D}$:

$$\text{Network Training:}\ \min_\theta \frac{1}{n}\sum\nolimits_{i=1}^{n} L\left(f_\theta(x_i), y_i\right).$$

Generally, L is chosen such that the *training loss*, $\frac{1}{n}\sum_{i=1}^{n} L\left(f_\theta(x_i), y_i\right)$, is large when the network makes many classification errors and small when the network performs well. Moreover, L should have desirable mathematical properties such as smoothness, boundedness, and continuity.

The minimization described in the display above can then be performed using empirical algorithms such as stochastic gradient descent (SGD).[23] These algorithms need an initial set of parameters, θ_0, on which it makes iterative adjustments. One could choose θ_0 randomly: this is called training *from scratch*. However, a common practice is initializing θ using parameters of a network trained on a

23 Léon Bottou, On-Line Algorithms and Stochastic Approximations, in: D. Saad (ed.), *Online Learning and Neural Networks*, Cambridge 1998; Léon Bottou/Olivier Bousquet, The Tradeoffs of Large Scale Learning, in: J. C. Platt et al. (eds.), *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2007, 161–168.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

different task and applying the optimization algorithm from the pretrained starting point. This practice, called *fine-tuning*, reduces the computational cost of training a network, and allows one to take advantage of image filters and feature extractors (created for a different task using a much larger dataset) that may not be achievable if one were to train from scratch on $\mathcal{D}$ (because n is too small).[24]

For example, today there are massive image datasets with millions and even billions of images for which network models have been trained; even if the images and descriptions in such existing trainings are quite different from those in our specific dataset. Under fine-tuning, we borrow an existing pipeline architecture and fully trained model to get our initialization $\mathcal{D}$, $u\theta_0$. One notable advantage of the deep net frameworks developed in Section 2.3 is that they are amenable to fine-tuning despite having undergone significant structural changes from the original setting.

Hence, we see that training requires only the specification of descriptions, prediction rule $\hat{y}$, and loss L to represent the task of interest. We have not yet fully described how the loss and prediction functions are chosen. This depends on details of the allowable descriptions of images: they could be fixed phrases chosen from a list; they could be combinations of such phrases; there could even be a grammar of allowable phrase combinations.

Section 2.1 Single-label Classification

The *single-label* classification problem has C different possible descriptions, and we must assign one of those C descriptions to each image. The collection of all images with the same description is called a *class*. Most modern deep neural networks were originally developed for classifying images in the benchmark ImageNet dataset[25] with $C = 1000$ classes (fish, bird, tree, etc.). In the context of works of art, the classes might represent the artists; the period; or a description of the objects represented in the scene. Most prior work for fine art classification solves single label problems. (See Section 4.)

Formally, for each image x, its description can be identified with its class's serial number, so the target can be represented as

$$y \in \{1, \ldots, C\},$$

and we call y the *true class* to which the image belongs. The feature vector $f_\theta(x)$ contains C numbers $f_\theta \equiv \left(f_\theta^c\right)_{c=1}^{C}$ with larger numbers meaning "more likely" and

24 Jason Yosinski et al., How Transferable Are Features in Deep Neural Networks?, in: Z. Ghahramani et al. (eds.), *Advances in Neural Information Processing Systems* 27, 2014, 3320–3328.

25 Jia Deng at al., Imagenet: A Large-Scale Hierarchical Image Database, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, 248–255.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

smaller numbers "less likely." The decision of the network chooses the class (i.e. the description) having the highest score: In other words, $\hat{y}$ is the class corresponding to the largest score:

$$\hat{y}(x) = \operatorname{argmax}_{c \in \{1,\ldots,C\}} f_\theta^c(x).$$

For the single-label classification task, modern machine learning practice typically employs the *cross-entropy* (CE) loss:

$$L_{CE}(f_\theta(x), y) \equiv -\log\left(\frac{\exp\left(f_\theta^y(x)\right)}{\sum_{c=1}^{C}\exp\left(f_\theta^c(x)\right)}\right).$$

Observe that L is small when the score of the correct class is much larger than the score of the incorrect classes. Additionally, note that the argument of the logarithm is always between 0 and 1. This argument, called the *softmax probability*, will be denoted by the function

$$p(s, y) \equiv \frac{\exp\left(s^y\right)}{\sum_{c=1}^{C}\exp\left(s^c\right)}.$$

Section 2.2 Simple Multilabel Classification

In the multilabel setting, *multiple* labels from a collection of C potential labels may have been used on any particular image, and one tasks the network to identify all labels simultaneously. For example, eligible labels might include "with water" and "with bridges," and we might want both labels identified when they are both present in an image. In contrast, some images might correctly be described only as "with bridges" or only as "with water."

In this setting, the true labels identify *subsets* $y \subseteq \{1, \ldots, C\}$. Such a subset y is typically represented as $y \in \{0, 1\}^C$ i.e. a length-C binary vector such that the c-th element, denoted y, is 1 or 0 according to whether the c-th label applies or not, respectively.

The network feature extractor, f_θ, remains the same, but the network's prediction,

$$\hat{y}(x) \equiv \left(\hat{y}^c(x)\right)_{c=1}^{C},$$

is now a binary vector with components

$$\hat{y}^c(x) = \mathbf{1}\left\{p(f_\theta(x), c) > \gamma\right\};$$

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Here, p is again the softmax probability, $\gamma \in [0, 1]$ is a *threshold parameter* chosen by the network engineer beforehand, and $\mathbf{1}\{\cdot\}$ is 1 or 0 depending on whether its argument is true. In Section 3.3, we discuss the choice of γ.

To train such a network, the CE loss that we discussed in the single label setting possesses a natural extension to the multilabel case called the *binary cross-entropy (BCE) loss:*

$$L_{BCE}\left(f_\theta\left(x\right), y\right) = -\left[\sum_{c=1}^{C} y^c \log p\left(f_\theta\left(x\right), c\right) + \sum_{c=1}^{C}\left(1 - y^c\right) \log\left(1 - p\left(f_\theta\left(x\right), c\right)\right)\right]$$

Training in the multilabel setting with the BCE loss is the machine learning community standard.

Section 2.3 Hierarchical Multilabel Classification

Multilabel classification problems sometimes possess an additional layer of structure that we call *hierarchical multilabel classification*: this is the case with the FARL classification system. In this problem, experts will never combine multiple labels in a completely arbitrary way; for example, it makes no sense to label the subject of an image as both "head to left" and "head to right." In fact, the possible descriptions in the FARL dataset can be identified with the nodes of a hierarchical structure.

Abstractly, the Frick classification system can be considered domain-specific language consisting of a set, $\mathcal{T}$, of C different *terms*—subsets of which can be grouped together into *phrases*. Yet, the grouping of terms into phrases must follow a syntax i.e. only certain combinations of terms in certain orders are allowed. Mathematically, we can represent phrases as ordered tuples of terms. Letting $\mathcal{P}$ represent the collection of all legal phrases, their relationship can be represented as

$$\mathcal{P} \subset \left\{\left(t_1, t_2, \ldots\right) : t_i \in \mathcal{T}\ \forall\ i\right\}.$$

In practical classification systems, the collection of all phrases has a hierarchical structure, where the topmost elements (the leftmost in the tuple) are the terms that can function as phrases all on their own, and the lower phrases in the hierarchy are syntactically legal continuations of higher phrases in the hierarchy.

In the context of FARL's classification system, components such as "portraits," "men," and "with hands" are examples of terms; full headings such as "Portraits: Men: With hands" and "Landscapes: With water" are examples of phrases; these phrases would be encoded as the tuples (portrait, men, with hands) and (landscapes, with water), respectively.

In the hierarchical multilabel setting the targets are then $y \in \mathcal{P}$. To incorporate the classification system's syntax into our prediction and loss functions, we first define the concepts of prefixes and syntactical continuations.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Definition (Prefix). If $r = (t_1, \ldots, t_l)$ is a valid phrase consisting of all l terms, then each tuple $r_k = (t_1, \ldots, t_k)$ where $1 \leq k \leq l$ is a *prefix* of r. We let $\mathcal{P}_+$ denote the set of all prefixes of phrases in $\mathcal{P}$.

Observe that the full-phrases are prefixes themselves, i.e., $\mathcal{P} \subseteq \mathcal{P}_+$.

Definition (Syntactical continuation function). For a set of phrases, $\mathcal{P}$, built from terms, $\mathcal{T}$, the *syntactical continuation* function, $\mathcal{S}: \mathcal{P}_+ \to 2^{\mathcal{T}}$, lists all terms that may follow the prefix i.e.

$$\mathcal{S}(r) = \left\{t \in \mathcal{T}: \ (r, t) \in \mathcal{P}_+\right\},$$

where (r, t) denotes the tuple created by appending the term, t, to the tuple r.

Since $\mathcal{T}$ has C elements and the score vector $f_\theta(x)$ has C components, we can assign a one-to-one correspondence between them—denoting $f_\theta^t(x)$ as the component of $f_\theta(x)$ associated with term t. Then, for a target phrase y of length L, define the *hierarchical cross-entropy* (HCE) loss:

$$L_{HCE}(f_\theta(x), y) = -\sum_{\ell=0}^{L} \sum_{t \in \mathcal{S}(y^{1:\ell})} \left[\mathbf{1}\{y^{\ell+1} = t\} \log\left(p_{y^{1:\ell}}(f_\theta(x), t)\right) + \mathbf{1}\{y^{\ell+1} \neq t\} \log\left(1 - p_{y^{1:\ell}}(f_\theta(x), t)\right)\right]$$

where we consider a *branch-specific softmax,*

$$p_{y^{1:\ell}}(f_\theta(x), t) \equiv \frac{\exp\left(f_\theta^t\right)}{\sum_{t' \in \mathcal{S}(y^{1:\ell})} \exp\left(f_\theta^{t'}(x)\right)},$$

and we follow the conventions that

$$\mathbf{1}\left\{y^{L+1} = t\right\} = \left(1 - \mathbf{1}\left\{y^{L+1} \neq t\right\}\right) = 0,$$

and that $\mathcal{S}\left(y^{1:0}\right)$ is the empty set.

Given the scores $f_\theta(x)$ and a preset threshold, γ, the *syntax-aware classifier*, $d_\gamma(f_\theta(x))$, is the procedure described in Algorithm 1.

Algorithm 1:

INPUT: $f_\theta(x)$, γ

$\hat{r} = (); \ell = 0$

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

WHILE $\left(\mathcal{S}\left(\hat{r}\right) \neq \varnothing \textbf{ AND } \min_{t \in \mathcal{S}(\hat{r})} p_{\hat{r}}\left(f_\theta\left(x\right), t\right) > \gamma\right)$:

$$\hat{r} \leftarrow \left(\hat{r}, \arg\max_{t \in \mathcal{S}(\hat{r})} p_{\hat{r}}\left(f_\theta\left(x\right), t\right)\right)$$

OUTPUT: $\hat{r}$

END

For an arbitrary image, x, the network prediction is, the network prediction is $\hat{y}_\theta\left(x\right) = d_\gamma\left(f_\theta\left(x\right)\right)$.

Section 3. Empirical Results

To demonstrate the advantage of the hierarchical classification framework using the HCE loss and syntax-aware classifier, we show experiments comparing three approaches for classifying the FARL dataset corresponding to different approaches described in Section 2:

1. **(SL):** Single-class classification, where we consider each of the 642 unique FARL headings to be a unique class.
2. **(SML):** Simple-multilabel classification, where any image can be given any subset of the 378 single-term labels that comprise headings in the FARL system.
3. **(HML):** Hierarchical multilabel classification trained with decisions made through the syntax-aware classifier. In this context, the FARL classification contains 642 different phrases built from 378 terms.

In these experiments, we compare the performance of each method on a *test* dataset that contains a *different* set of images than the training data ($\mathcal{D}$). Thus, the training set serves to train the classifiers, while the testing set evaluates on the performance of the trained classifiers on new, previously unseen data.

Section 3.1 Experimental Details

Specifically, we use a dataset of 57,803 images from the American Portraits Collection within the FARL Photoarchive, which have already been labeled by photoarchivists according to the FARL classification system. This dataset is randomly split into a training subset, consisting of 80% of the images (46,242 images), and a testing subset, containing the remainder (11,561 images). Adding to the challenge of the task, different terms and phrases possess varying numbers of training examples. For instance, the most populous term, "portraits," possess 17,075 training examples while the term "with sheep and goats" possesses only one example in the training set. Similarly, the most populous phrase, "genre," possess 1,715 examples

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

while the "animals: cattle: with figures" possesses only one example. See Section 3.5 on how this class imbalance affects performance.

The images range in size from 55KB (1213x1536 pixels) to 4.4MB (4868x2856 pixels). Pixels are standardized by subtracting the (red, green, or blue) channel mean and dividing by the (red, green, or blue) channel standard deviation. We follow the same data augmentation steps typically performed on the ImageNet dataset: Images are rescaled such that the smaller dimension is 256 pixels, and then a random crop (in training) or a center crop (in testing) of 224x224 pixels is extracted. We do *not* apply random horizontal flips—as is common in deep networks trained on computer vision applications—since the FARL headings distinguish between portraits facing left or right.

In all settings, we train the feature extractor (f_θ) by fine-tuning all the parameters of a ResNet152 architecture pretrained on ImageNet. The pretrained model was downloaded directly from the PyTorch ModelZoo.[26] Following common practice, we minimize the task-specific loss using stochastic gradient descent (SGD), with a momentum of 0.9, and a weight decay of 10^{-4}. Our networks are trained on 4 GPUs, with a total batch size of 32 images, for 100 epochs. The initial learning rate is annealed by a factor of 10 at epoch 34 and 67. We train models with initial learning rates 0.1 and 0.01—picking the one resulting in the best true positive rate in the last epoch. In all three tasks, initial learning rate 0.01 produced the best model.

Section 3.2 Performance Metrics

In order to compare the performance of the three methods on the test set, we gather the following statistics for each of the terms in the FARL classification system (i.e. elements of $\mathcal{T}$ defined in Section 2):

- **Term Count (n_i'):** Number of images such that the i-th term applies.
- **True Positives (TP_i):** Number of images such that the i-th term applies, and the network correctly labeled them with that term.
- **False Negatives (FN_i):** Number of images such that the i-th term applies, but the network didn't label them with that term.
- **True Negatives (TN_i):** Number of images such that the i-th term doesn't apply, and the network correctly didn't label them with that term.
- **False Positives (FP_i):** Number of images such that the i-th term doesn't apply, but the network incorrectly labeled them with that term.

26 PyTorch ModelZoo, URL: https://pytorch.org/docs/stable/model_zoo.html [last accessed: April 2, 2021].

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Observe that the following relationships apply for each i:

$$\begin{aligned} TP_i + FN_i + TN_i + FN_i &= n', \\ TP_i + FN_i &= n_i', \\ TN_i + FP_i &= n' - n_i', \end{aligned}$$

where n' is the total number of images in the testing set.

Following common practice, to measure model performance on a particular term, we will use the *precision*, *recall*, and *F1 scores* defined, respectively, as follows:

$$Prec_i \equiv \frac{TP_i}{TP_i + FP_i}, Rec_i \equiv \frac{TP_i}{TP_i + FN_i}, F_1^i \equiv 2\frac{Prec_i \times Rec_i}{Prec_i + Rec_i}.$$

Roughly, precision measures the network's ability to ignore non-examples of a term, while recall (sometimes called true positive rate) measures the network's ability to identify true examples of a term. The F1 score aggregates precision and recall into an overall measure of performance.[27]

To evaluate performance over the *entire testing set*, we use a weighted aggregate of these scores:

$$Rec \equiv \sum_{i=1}^{|\mathcal{T}|} w_i Rec_i, Prec \equiv \sum_{i=1}^{|\mathcal{T}|} w_i Prec_i, F_1 \equiv \sum_{i=1}^{|\mathcal{T}|} w_i F_1^i,$$

where $|\mathcal{T}| = 378$ is the number of terms in the classification system, and the weights,

$$w_i \equiv \frac{n_i'}{\sum_{i=1}^{|\mathcal{T}|} n_i'},$$

capture the relative frequencies of instances of terms in the testing set.

Section 3.3 Determination of Threshold

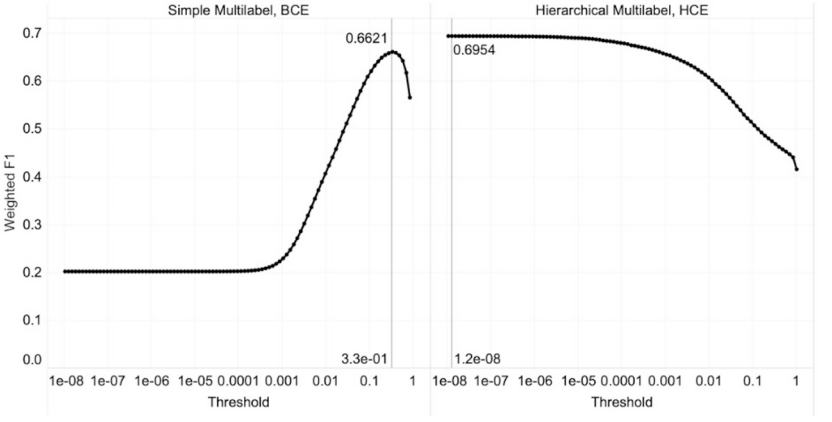
Recall from Section 2 that both the SML and HML settings require a pre-determined choice for the decision threshold, γ. In practice, we determine this by measuring the final performance on test data for each candidate choice and choosing the best candidate value. In Fig 1.7, we plot an experiment showing the weighted-F1 scores resulting from varying this threshold. The difference in the order of optimal-threshold magnitudes for the SML setting compared to the HML setting reflects the difference between the uniform nature of the BCE loss compared to the adaptive

27 For a more detailed discussion, see: Koo Ping Shung, Accuracy, Precision, Recall or F1? (2018), in: *Towards Data Science. Medium*, URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [last accessed: April 2, 2021].

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

structure of the HCE loss. Using the best threshold from Fig 1.7 in the remainder of this section, we can then compare the different methods.

Fig 1.7: Weighted-F1 scores versus prediction threshold. Each row corresponds to a different classification approach (described in Section 2) used on images from the Frick Art Reference Library's Photoarchive. In each array cell, we plot (on the logarithmically spaced x-axis) 100 prediction thresholds (γ) and (on the y-axis) their resulting weighted-F1 score on the testing dataset. The largest score and its corresponding threshold are annotated on the top and bottom of each plot, respectively. Training details are given in Section 3.1. Courtesy of the authors.



Section 3.4 Comparison of Performance

In Table 1, for each of the settings described in Section 4.1, we report weighted precision, recall, and F1 scores over the entire testing set. Using the SL setting as a control, we see that changing to the SML setting improves recall (i.e. reduces false negatives) and overall performance but decreases precision (i.e. increases false positives). In comparison, the HML setting improves all three performance metrics compared to both the SL and SML settings.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Table 1. Performance Metrics on Testing Data. Each row corresponds to a different classification approach (described in Section 2) used on images from the Frick Art Reference Library Photoarchive. The first column shows the best threshold chosen according to the weighted-F1 score on the testing data. The last three columns show the weighted precision, recall, and F1 score on the testing data (described in Section 3.2). Training details are given in Section 3.1.

Setting	Best Threshold	Precision	Recall	F1 Score
SL-CE	N/A	0.6538	0.6663	0.6577
SML-BCE	0.33	0.6364	0.7018	0.6621
HML-HCE	1e-08	0.6758	0.7226	0.6954

These results are intuitive. The greater flexibility afforded by predicting single terms out of multiple-term phrases, in the multilabel regime, rather than predicting the complete multi-term phrase, in the single label regime, allows the network to better identify positive instances of labels (better recall) in both the SML and HML case. In the HML case, this performance is enhanced even further by the explicit incorporation of the classification domain-specific language into the training loss—allowing for a more focused and efficient tuning of parameters.

On the other hand, the increased flexibility results in a larger space of possible outputs, which, in turn, increases the possibilities for false positives. For example, in the SML case, all combinations of terms are considered syntactically valid results for the network—even those not in the FARL classification system itself. We see from the decreased precision that the SML case is indeed hurt by failing to use syntax constraints. In contrast, the HML avoids this flaw by reducing the space of possible outputs to the most economical representation: *the only predictions made by the syntax-aware classifier are those in the FARL classification system*. This combined with the advantage of a loss and predictor specialized to the FARL system likely led to HML's increase in precision.

Overall, these results demonstrate the advantage of incorporating the syntax of the classification system into the neural network itself. These performance gains come as a direct consequence of the close collaborative efforts of the AI researchers and art historians where the latter group provided important insights into the structure of the dataset as well as practical aspects of which guided the development of the syntax-aware classifier and HCE loss (Section 2.3).

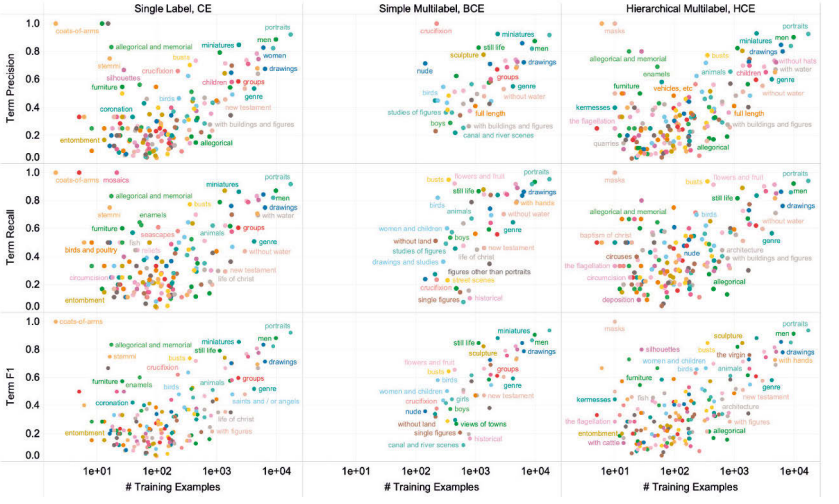
Section 3.5 Impact of Sample Size

Another prominent factor determining network performance is the number of examples within each class in the training data. For individual terms in the testing

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

set, Fig 1.8 shows that performance as measured by precision, recall, and F1 score are all approximately proportional to the logarithm of the number of examples of the term in the training dataset. Yet variance is still high even for terms with roughly equal numbers of training examples. Nonetheless, the trends in Fig 1.8 provide an approximate guideline for the necessary amount of prerequisite training data needed to reach a reasonable performance, which is a useful reference for fine art institutions who wish to adopt deep learning methods in the future.

Fig 1.8: Sample size versus precision and recall. Each row corresponds to a different classification approach (described in Section 2) used on images from the Frick Art Reference Library's Photoarchive. The optimal thresholds in the multilabel settings are chosen according to the experiment in Fig 7. Each array row corresponds to a term-wise performance described in Section 3.2. In each array cell, we plot a scatterplot of terms with non-zero true-positive counts: the x-axis (logarithmically spaced) shows the number of examples of a term in the training set, and the y-axis shows the trained network's performance on that term. A subsample of the points is labeled, in the same color, with their corresponding term. Training details are given in Section 3.1. Courtesy of the authors.



This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Section 4. Related Works

The idea of leveraging computer vision techniques to classify fine art images predates the popularity of deep learning. From the early 2000s to the mid-2010s, researchers tried to automate the annotation of fine art images, usually paintings, by applying regression and classification methods on handcrafted features emerging from the wavelet transform,[28] the scale-invariant feature transform (SIFT),[29] the GIST descriptor,[30] and the histogram of oriented gradients descriptor (HOG).[31] These works predominantly focused on single-label classification tasks, such as the identification of the creator of an artwork. Agarwal et al.—who use five such features for identifying the artist and genre of artworks—provide a comprehensive survey of such works.[32]

The rise of deep learning has led to a paradigm shift. Instead of training a classifier on handcrafted features, deep neural networks are trained *both* to extract the relevant features from an image (in the earlier convolutional layers) and to make the classification decision (in the later linear layers). As a result, around 2015—inspired by the human-level performance of deep convolutional neural networks (CNNs)—researchers shifted their focus to the automatic classification of artworks by fine-tuning CNNs.

The most commonly utilized dataset for this later line of work is WikiArt—a growing online collection of approximately 80,000 digital images of fine art paintings.[33] It includes four tasks: artist identification, style identification, genre identification, and time period identification. For example, Cetinic et al.,[34] Saleh et

28 Jia Li/James Ze Wang, Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models, in: *IEEE Transactions on Image Processing* 13 (3/2004), 340–353.

29 Fahad Shahbaz Khan/Joost Van de Weijer/Maria Vanrell, Who Painted this Painting?, in: *2010 CREATE Conference*, 2010, 329–333.

30 Matthijs Douze et al., Evaluation of Gist Descriptors for Web-Scale Image Search, in: *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, 1–8.

31 Xiaoyu Wang/Tony X. Han/Shuicheng Yan, An HOG-LBP Human Detector with Partial Occlusion Handling, in: *2009 IEEE 12th International Conference on Computer Vision*, Kyoto 2009, 32–39.

32 Siddharth Agarwal et al., Genre and Style Based Painting Classification, in: *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, 2015, 588–594.

33 WikiArt, URL: www.wikiart.org [last accessed: April 2, 2021].

34 Eva Cetinic /Tomislav Lipic/Sonja Grgic, Fine-Tuning Convolutional Neural Networks for Fine Art Classification, in: *Expert Systems with Applications* 114 (2018), 107–118.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

al.,[35] Tan et al.,[36] Hentschel et al.,[37] and Lecoutre et al.[38] propose different algorithms for tackling the style identification task based upon fine-tuning CNNs and all show how such algorithms achieve state-of-the-art results. The first three of the list of five papers further demonstrate the ability of CNNs to perform well in the artist and genre identification tasks. These advances address various aspects of network architecture design, initialization, and sample size efficiency, but all within the single-label classification regime. Cetinic et al. provides a comprehensive summary and comparison of these works.

Another notable line of research, originating in the mid-2010s, is that inspired by a team from the Rijksmuseum in Amsterdam, the Netherlands. In 2014, Mensink and Gemert released a dataset of 112,039 images of fine art from the collection of the Rijksmuseum.[39] They identified four tasks on this dataset (collectively called the Rijksmuseum Challenge): artist attribution (single-label classification); art-type identification (simple-multilabel classification); materials identification (simple-multilabel classification); and creation year prediction (regression). They then created benchmark performance metrics obtained by encoding the images as Fisher vectors followed by regression or max-margin classification.

In 2017, Strezoski and Worring extended the Rijksmuseum dataset into the OmniArt dataset, which has 432,217 images, adding images and metadata from the Rijksmuseum's collection as well as additional Open Access images from the holdings of the Metropolitan Museum of Art (the Met).[40] They established new benchmark results on the four Rijksmuseum Challenge tasks using fine-tuned CNNs. In 2018, Strezoski and Worring[41] expanded OmniArt to contain more than two

35 Babak Saleh/Ahmed Elgammal, Large-Scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature, in: *International Journal for Digital Art History* (2/2016), 71–93.

36 Wei Ren Tan et al., *Ceci n'est pas une pipe*: A Deep Convolutional Network for Fine-Art Paintings Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3703–3707.

37 Christian Hentschel/Timur Pratama Wiradarma/Harald Sack, Fine Tuning CNNS with Scarce Training Data—Adapting ImageNet to Art Epoch Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3693–3697.

38 Adrian Lecoutre/Benjamin Negrevergne/Florian Yger, Recognizing Art Style Automatically in Painting with Deep Learning, in: Min-Ling Zhang/Yung-Kyun Noh (eds.), *Proceedings of the Ninth Asian Conference on Machine Learning*, *PMLR* 77, 2017, 327–342.

39 Thomas Mensink/Jan Van Gemert, The Rijksmuseum Challenge: Museum-Centered Visual Recognition, in: *ICMR '14: Proceedings of International Conference on Multimedia Retrieval*, Glasgow 2014, 451–454.

40 Gjorgji Strezoski/Marcel Worring, OmniArt: Multi-Task Deep Learning for Artistic Data Analysis (2017), in: *arXiv* [preprint], URL: arXiv:1708.00684 [last accessed: April 2, 2021].

41 Gjorgji Strezoski/Marcel Worring, OmniArt: A Large-Scale Artistic Benchmark, in: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14 (4/2018), 1–21.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

million images by including Open Access images from additional museums. They added additional metadata that is amenable to the single-label, simple-multilabel, and regression tasks.

In our work with the FARL dataset, we also employed the fine-tuning technique on CNNs. Yet most prior work, such as that by Cetinic et al. and Strezoski and Worring, tackles single-label classification or simple-multilabel classification. In contrast, FARL's in-house classification system is a much better fit to the hierarchical multilabel classification problem. Therefore, our project is distinct from these precedents in two ways. First, unlike in the single-label setting—which requires one to pick only the network prediction with the *highest score*—the multilabel setting requires one to choose all classes whose network-predicted scores exceed some threshold. (See Section 2.) Thus, additional attention must be given to the determination of the threshold. (See Section 3.) Second, fine-tuning a pretrained, single-label network for another single-label or simple-multilabel classification task requires replacing only the last linear layer of the network with another linear layer with dimensions matching the new problem. In contrast, to adapt to FARL's *hierarchical* labeling system, we engineered both a novel *hierarchical cross-entropy loss* as well as a novel *syntax-aware classifier* (described in Section 2), which, as we demonstrated in Section 3, results in a higher performing network.

A related labeling system is Iconclass,[42] which is also a hierarchical classification system for images of fine art but is structurally different from FARL's system. Unlike Iconclass, headings in FARL's system are comprised of *common components*: each multi-term phrase in the hierarchy is composed from a shared set of component terms, whereas the headings in the Iconclass hierarchy are predominantly unique descriptors. (See Section 1.2 and 2.) Additionally, Iconclass, with 28,000 total headings and growing,[43] is much larger than FARL's system. Iconclass may be amenable to the hierarchical multilabel classification framework, but a significant amount of semantic pre-processing of the headings would be required.

Within the hierarchical setting, a prior work by Belhi et al. uses deep learning on the hierarchical multilabel problem of the WikiArt, the Met, and Rijksmuseum datasets for a *two-level* hierarchy.[44] This hierarchy consists of (i) a general asset-type (for example, pottery, paintings, etc.) and (ii) specific characteristics for each asset-type (for example, predicting artist and style for asset-type "paintings"). The

42 Leendert D. Couprie, Iconclass: An Iconographic Classification System, in: *Art Libraries Journal* 8 (2/1983), 32–49.

43 Iconclass, URL: www.iconclass.org [last accessed: April 2, 2021].

44 Abdelhak Belhi/Abdelaziz Bouras/Sebti Foufou, Towards a Hierarchical Multitask Classification Framework for Cultural Heritage, in: *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, Aqaba 2018, 1–7.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

authors tackled this problem by training two types of deep networks: one single-label classification network for predicting the asset-type, and another asset-specific multitask CNN[45] for each asset-type.

In contrast to the two-level hierarchy of Belhi et al., FARL's hierarchy could sometimes be five or six levels deep with hundreds of unique non-terminal[46] and terminal[47] branching points—many of which are relevant to only a handful of images. Thus, applying the algorithm of Belhi et al. to the FARL system would involve (i) training new CNNs at each non-terminal branching point and (ii) training a multitask CNN at each terminal point. Due to the depth of the hierarchy, hundreds of models would have to be trained. Not only are the computational costs for training hundreds of networks unattainable for most institutions, but also such an approach would train the parameters of each component network on a *subset* of the data, which would likely cause deterioration in performance. (See Section 3.4.) In contrast, the method described in Section 2 provides a solution that can predict all headings in the FARL hierarchy by training only *one* network on the *entire* prelabeled dataset.

Outside the field of fine art classification, researchers in text and image annotation as well as protein identification have extensively studied the hierarchical classification problem and, for the most part, relied on hand-engineered features that are input into decision-tree or max-margin classifiers. Nakano, Cerri, and Vens provide a thorough survey with insightful discussions.[48] Methods leveraging deep learning appeared only recently. Of particular interest are the works by Wehrmann et al.[49] and by Wehrmann, Cerri, and Barros,[50] which propose various convolutional and recurrent neural networks, respectively, for the hierarchical classification problem. Their approach relies on architectures markedly different than those pretrained for single-label classification. Thus, since most open source pretrained computer vision models were trained for the single-label task—usually on

45 Following a shared series of feature extraction layers, a multitask CNN branches into parallel linear layers that each performs a single-label classification or regression task such as artist or genre prediction.

46 A non-terminal branch point is a location in the hierarchy, associated with some prefix, such that there exists a next-level-down term that, once appended, will create another prefix that is not a full phrase.

47 A terminal branch point is a location in the hierarchy, associated with a prefix, such that appending any next-level-down term to the prefix will create a full syntactically valid phrase.

48 Felipe Kenji Nakano/Ricardo Cerri/Celine Vens, Active Learning for Hierarchical Multi-Label Classification, in: *Data Mining and Knowledge Discovery* 34 (5/2020), 1496–1530.

49 Jônatas Wehrmann et al., Hierarchical Multi-Label Classification with Chained Neural Networks, in: *SAC'17: Proceedings of the Symposium on Applied Computing*, Marrakech 2017, 790–795.

50 Jônatas Wehrmann/Ricardo Cerri/Rodrigo C. Barros, Hierarchical Multi-Label Classification Networks, in: *Proceedings of the 35th International Conference on Machine Learning*, 2018, 5075–5084.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

ImageNet—their approach, unlike ours, cannot leverage transfer learning, which improves performance significantly in the low-sample size regime.

During the preparation of this chapter, we learned of a pilot project launched by Lincoln et al. at the Carnegie Mellon University (CMU) Libraries to tackle the tasks of visual similarity search, duplicate and close-match identification, and streamlining image tagging within CMU's General Photograph Collection (GPC).[51] This project resulted in a deep learning pipeline for these tasks that was further refined through testing by and feedback from CMU Libraries staff. In their white paper, Lincoln et al. discuss how computer vision might be used within Photoarchives to streamline cataloging and improve searching. Furthermore, the authors emphasize the importance of using a human-in-the-loop process to tackle these tasks to prevent error and correct any biases promoted by the training set. We strongly agree with these conclusions.

Yet, although the CMU team's tagging task appears similar to that introduced in this chapter, it relies on different methodologies. Our approach consists of training a network (starting from a network pretrained on ImageNet) to *directly predict* the heading associated with an unlabeled image while the CMU team's approach is to identify unlabeled images (by using an ImageNet-pretrained network that is not further trained on the GPC) similar to a particular labeled "seed" image, which the human editors must identify.

Our work and that of the CMU team require human experts to validate the accuracy of the suggestions at the end of the pipeline. As described by Lincoln et al., their pipeline requires that editors locate the seed images and, occasionally, it returns inconveniently large subsets of similar images. In contrast, our Zooniverse app directly presents images—one by one—with a predicted label and the human expert can choose "correct" by swiping right or "incorrect" by swiping left or input an alternative label. Thus, our direct-prediction approach to vetting results is significantly faster than that proposed by the CMU team. The white paper published by Lincoln et al. does not present the performance metrics of their pipeline, so a comparison of accuracy cannot yet be determined.

Section 5. Conclusion

In this chapter, we establish the considerable benefits of collaborations between AI researchers and cultural heritage preservationists. We also demonstrate that deep neural networks can be adapted to classify images according to specialized hierarchical, multilabel classification systems. This adaptation requires only simple

51 Matthew Lincoln et al., CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative (2020), in: Carnegie Mellon University. Preprint, doi:10.1184/R1/12791807.v2.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

modifications to the network's loss function and prediction rule, thus leaving the feature extractor unchanged and allowing for the utilization of pretrained models, through fine-tuning, to achieve better performance. Finally, we provide original experiments on digitized images in FARL's Photoarchive, thus indicating the validity and potential of this approach. As discussed above, we achieved a notable improvement in performance by incorporating the FARL classification system into the network through our proposed HCE loss and syntax-aware classifier.

These experiments further indicate that the accuracy for a given label depends logarithmically on the number of training images tagged with that label. For example, results in Section 3.4 suggest that roughly 100 training images are necessary to achieve 30%–50% success in canonical performance metrics, 1,000 images for 50%–80%, and 10,000 images for more than 80%. These experimental insights provide a promising point of reference for future partnerships between art historians and computer scientists.

As noted above, we anticipate that this technology will be ready for deployment in 2022 and once introduced, it is certain to streamline Library staff workflow by relieving photoarchivists of a time-consuming aspect of cataloguing. Thus, for the photoarchivists, the successful partnership as documented in this chapter demonstrates a clear advantage. For the computer vision researchers, the Photoarchive's in-house classification system motivates useful investigations in engineering the syntax of hierarchical, domain-specific languages into neural networks.

Bibliography

AGARWAL, Siddharth/KARNICK, Harish/PANT, Nirmal/PATEL, Urvesh, Genre and Style Based Painting Classification, in: *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, 2015, 588-594.

BALDI, Pierre/SADOWSKI, Peter/WHITESON, Daniel, Searching for Exotic Particles in High-Energy Physics with Deep Learning, in: *Nature Communications* 5 (1/2014), 1–9.

BELHI, Abdelhak/BOURAS, Abdelaziz/FOUFOU, Sebti, Towards a Hierarchical Multitask Classification Framework for Cultural Heritage, in: *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, Aqaba 2018, 1–7.

BOTTOU, Léon/BOUSQUET, Olivier, The Tradeoffs of Large Scale Learning, in: J.C. PLATT et al. (eds.), *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2007, 161–168.

BOTTOU, Léon, On-Line Algorithms and Stochastic Approximations, in: D. SAAD (ed.), *Online Learning and Neural Networks*, Cambridge 1998.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Cetinic, Eva/Lipic, Tomislav/Grgic, Sonja, Fine-Tuning Convolutional Neural Networks for Fine Art Classification, in: *Expert Systems with Applications* 114 (2018), 107–118.

Couprie, Leendert D., Iconclass: An Iconographic Classification System, in: *Art Libraries Journal* 8 (2/1983), 32–49.

Deng, Jia/Dong, Wei/Socher, Richard/Li, Li-Jia/Li, Kai/Li, Fei-Fei, Imagenet: A Large-Scale Hierarchical Image Database, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, 248–255.

Douze, Matthijs/Jégou, Hervé/Sandhawalia, Harsimrat/Amsaleg, Laurent/Schmid, Cordelia, Evaluation of Gist Descriptors for Web-Scale Image Search, in: *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, 1–8.

Goodfellow, Ian/Bengio, Yoshua/Courville, Aaron, *Deep Learning*, Cambridge 2016.

He, Kaiming/Zhang, Xiangyu/Ren, Shaoqing/Sun, Jian, Deep Residual Learning for Image Recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas 2016, 770–78.

Hentschel, Christian/Wiradarma, Timur Pratama/Sack, Harald, Fine Tuning CNNS with Scarce Training Data—Adapting ImageNet to Art Epoch Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3693–3697.

Khan, Fahad Shahbaz/Van de Weijer, Joost/Vanrell, Maria, Who Painted this Painting?, in: *2010 CREATE Conference*, 2010, 329–333.

Knox, Katharine McCook, *The Story of the Frick Art Reference Library: The Early Years*, New York 1979.

Lecoutre, Adrian/Negrevergne, Benjamin/Yger, Florian, Recognizing Art Style Automatically in Painting with Deep Learning, in: Min-Ling Zhang/Yung-Kyun Noh (eds.), *Proceedings of the Ninth Asian Conference on Machine Learning, PMLR* 77, 2017, 327–342.

Li, Jia/Wang, James Ze, Studying Digital Imagery of Ancient Paintings by Mixtures of Stochastic Models, in: *IEEE Transactions on Image Processing* 13 (3/2004), 340–353.

Lincoln, Matthew, et al., CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative (2020), in: Carnegie Mellon University. Preprint, doi:10.1184/R1/12791807.v2.

Nakano, Felipe Kenji/Cerri, Ricardo/Vens, Celine, Active Learning for Hierarchical Multi-Label Classification, in: *Data Mining and Knowledge Discovery* 34 (5/2020), 1496–1530.

Mensink, Thomas/Van Gemert, Jan, The Rijksmuseum Challenge: Museum-Centered Visual Recognition, in: *ICMR '14: Proceedings of International Conference on Multimedia Retrieval*, Glasgow 2014, 451–454.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms

Prokop, Ellen, Digital Art History for the Masses? The Role of the Public Digital Art History Lab, in: *Život umjetnosti: Journal for Modern and Contemporary Art and Architecture* 105 (2/2019), 196–213.

Saleh, Babak/Elgammal, Ahmed, Large-Scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature, in: *International Journal for Digital Art History* (2/2016), 71–93.

Sanger, Martha Frick Symington, *Henry Clay Frick: An Intimate Portrait*, New York 1998.

Shung, Koo Ping, Accuracy, Precision, Recall or F1? (2018), in: *Towards Data Science. Medium*, URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [last accessed: April 2, 2021].

Strezoski, Gjorgji/Worring, Marcel, OmniArt: A Large-Scale Artistic Benchmark, in: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14 (4/2018), 1–21.

Strezoski, Gjorgji/Worring, Marcel, OmniArt: Multi-Task Deep Learning for Artistic Data Analysis (2017), in: *arXiv* [preprint], URL: arXiv:1708.00684 [last accessed: April 2, 2021].

Tan, Wei Ren/Chan, Chee Seng/Aguirre, Hernán E./Tanaka, Kiyoshi, *Ceci n'est pas une pipe*: A Deep Convolutional Network for Fine-Art Paintings Classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, 3703–3707.

Yosinski, Jason/Clune, Jeff/Bengio, Yoshua/Lipson, Hod, How Transferable Are Features in Deep Neural Networks?, in: Z. Ghahramani et al. (eds.), *Advances in Neural Information Processing Systems* 27, 2014, 3320–3328.

Wang, Dayong/Khosla, Aditya/ Gargeya, Rishab/Irshad, Humayun/Beck, Andrew H., Deep Learning for Identifying Metastatic Breast Cancer (2016), in: *arXiv* [preprint], URL: arxiv:1606.05718 [last accessed: April 2, 2021].

Wang, Xiaoyu/Han, Tony X./Yan, Shuicheng, An HOG-LBP Human Detector with Partial Occlusion Handling, in: *2009 IEEE 12th International Conference on Computer Vision*, Kyoto 2009, 32–39.

Wehrmann, Jônatas/Cerri, Ricardo/Barros, Rodrigo C., Hierarchical Multi-Label Classification Networks, in: *Proceedings of the 35th International Conference on Machine Learning*, *PMLR* 80, 2018, 5075–5084.

Wehrmann, Jônatas/Barros, Rodrigo C./Dôres, Silvia N. das/Cerri, Ricardo, Hierarchical Multi-Label Classification with Chained Neural Networks, in: *SAC '17: Proceedings of the Symposium on Applied Computing*, Marrakech 2017, 790–795.

This content downloaded from 58.97.216.197 on Wed, 04 Sep 2024 03:20:28 UTC
All use subject to https://about.jstor.org/terms