Deep Mapping of the Visual Literature

Bill Howe
Information School
University of Washington

Po-shen Lee
Department of Electrical Engineering
University of Washington

Maxim Grechkin
Department of Computer Science & Engineering
University of Washington

Sean T. Yang
Department of Electrical Engineering
University of Washington

Jevin D. West
Information School
University of Washington

ABSTRACT

We consider how patterns of figure use in the scientific literature relate to impact, change over time, and vary across disciplines. We use a convolutional neural network to embed figures as feature vectors in a high-dimensional space, then visualize this space as a 2D heatmap to expose patterns. We consider how these patterns vary with respect to time, impact, and discipline, concluding that high-impact papers tend to include significantly more data-carrying figures (i.e., visualizations), despite a downward trend in such figures overall. We also show how this approach can be used to bootstrap targeted information extraction projects for specific figure types, describing one such project involving phylogenetic trees.

1. INTRODUCTION

Visualizations are the currency of scientific communication, but have largely been untouched by computational techniques due to the relative opacity of images compared to text or citations.

We analyze the use of visualization in the biomedical scientific literature by embedding figures as 2048-element vectors using a convolutional neural network, then inspecting and reasoning about the principal components of these vectors.

We find that photos and diagrams dominate the literature, but by visualizing the residuals from this baseline signal, we see an apparent overall increase in the use of complex diagrams and dense imagery.

However, among higher-impact papers, data visualizations are far more common than they are in lower-impact papers. When we consider differences by journal, we find that certain patterns of figures characterize journals, perhaps suggesting templates by through which authors can optimize the readability for a particular audience.

In this short paper, we describe the unsupervised learning pipeline we developed, as well as a set of online tools for interacting with the aggregate visual literature.

We consider this analysis to be an initial foray into understanding how the use of visualization affects scientific communication and impact, and how this use can be optimized. Moreover, these methods and tools allow identification of specific image types to enable specialized information extraction procedures. We will describe one such project, where we extract the data from phylogenetic trees in order to augment online databases with information from the literature.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053065

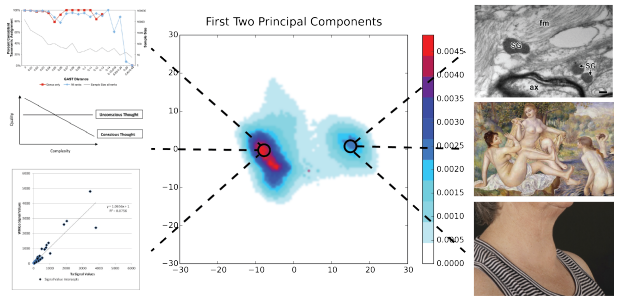

Figure 1: The first two primary components of 1.5M figures from PubMed embedded using the ResNet neural network architecture. The two obvious dense regions of the map roughly represent quantitative plots (left) and photographs (right). The red region represents higher densities of figures.

2. RELATED WORK

Computer vision techniques have been used in the context of conventional information retrieval tasks (retrieving papers based on keyword search), including some commercial systems such as D8taplex and Zanran. Search results from these proprietary systems have not been evaluated and do not appear to make significant use of the semantics of the images.

In 2001, Murphy et al. proposed a Structured Literature Image Finder (SLIF) system, targeting microscope images [16]. A decade later, Ahmed et al. [1, 2] improved the model for mining captioned figures. The latest version combines text-mining and image processing to extract structured information from biomedical literature. The algorithm first extracts images and their captions from papers, then clas-

sifies the images into six classes. Classification information and other metadata can be accessed via web service. However, SLIF focuses exclusively on microscropy images and does not extend to general figures.

Choudhury et al. [18] proposed a modular architecture to mine and analyze data-driven visualizations that included an extractor to separate figures, captions, and mentions from PDF documents [7] and a raw-data extractor for line charts [8]. Chen et al. [5] proposed their search engine named DiagramFlyer for data-driven figures that extracts axis labels and legend information via OCR and uses the extracted text to drive the search. Other studies have proposed informatics methods for retrieving maps of the brain through large-scale image and text mining on fMRI images [17].

Other projects explored the patterns of figure use in the literature. Hegarty et al. collected 1,133 articles from 9 psychology journals and found that articles with fewer graphs and more structural equation models were more frequently cited[13]. This result was not supported by other in different disciplines: Fawcett et al. studied the citations of 28,068 papers published in the top three journals specializing in ecology and evolution and found that heavy use of equations has a significant negative impact on citation rates[10]. Tartanus et al. reported a positive correlation between number of graphs and the impact factors in journal level by analyzing all papers published in 2010 from 21 selected journals in agriculture[19]. Other studies investigate how the use of figures differs by authorship patterns. Cabanac et al. analyzed 5,180 articles in the sciences and social sciences and found that *groups* of authors used significantly more tables and graphs than single authors[4]. Hartley et al. investigated approximately 2,000 articles from 200 journals in the sciences and social sciences. They found that men used 26% more figures than women, but found no significant difference in their use of tables. In addition, they didn't find significant differences between men and women in using either graphs and figures or tables in social science articles[11]. Since counting figures manually is extremely time-consuming, all of these studies were limited to specific domains on a relatively small number of papers and journals. Our approach is to automate the analysis using computer vision techniques and machine learning, scale it to a large corpus of papers to allow broader inferences, and release the software and labeled data for other researchers to use.

3. PROCESSING THE FIGURE CORPUS

We process a sample of 1.5 million figures from selected randomly from all papers uploaded to PubMed. In previous work [15], we worked with the entire set of eight million figures from PubMed and developed a supervised classification pipeline to label each figure into broad categories: diagram, plot, photograph, or table. We considered how the density of these categories related to impact, evolved over time, and differed between journals, finding that plots and diagrams are correlated with impact, suggesting the importance of visualization in scientific communication.

The limitation of that work was the need to provide class labels a priori, which cannot capture the variety and nuance of how researchers present information visually.

In this paper, we present preliminary results from an unsupervised learning approach, attempting to understand — qualitatively and quantitatively — how patterns in the visual literature relate to impact, change over time, and vary by field.

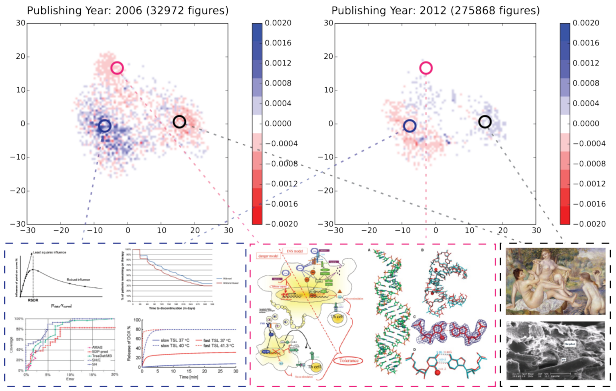

Figure 2: Density deviation by year. The average figure density for 2006 (left) and 2012 (right) are compared to the global average from 2003-2013. Red regions represent a positive deviation from the average and blue regions represent a negative deviation. We see a relative increase in photos and scientific illustrations and a relative decrease in quantitative plots.

We began by embedding each figure into a 2048-dimensional space using a pre-trained neural network, ResNet[12] using Keras [6], designed to extract features from a large corpus of natural images. We continue to explore a modified architecture and re-trained model specialized for artificial images in the scientific literature, but this model proved sufficient for preliminary results.

Deep Neural Networks (DNN) have received considerable attention over recent years in computer vision and object recognition [14]. Trained on millions of images by leveraging advances in GPU computing, these models can learn sophisticated representations of images without any manual feature engineering. Deep Residual Networks (ResNet) [12] is the winning model in ILSVRC and COCO 2015 competitions. In this work, we are using the ResNet with 50 layers that was pre-trained on ImageNet corpus [9] (a corpus of 1.2M natural images collected from the web). While the model was trained to recognize natural images, it is believed [20] that internal representation might be general, allowing for *transfer learning* - application of a model trained in one domain to another similar domain. In our case, we are taking are processing scientific images through ResNet model and obtaining a 2048 dimensional representation from a hidden layer, just beneath the final dense object recognition layer trained for ILSVRC competition. To visualize this hidden representation, we compute top two principal components of this 2048-dimensional space. Top principal components represent directions of largest variance in the space, allowing us to get meaningful visualization in an unsupervised way.

We then ran PCA on the 1.5M vectors and plotted the first two components as a heat map to convey density. The results are shown in Figure 1. There are two main regions visible in this 2D projection. On the left, the cluster roughly corresponds to plots, diagrams, and other figures featuring lines and text. On the right, the cluster roughly corresponds to photographs. The substructures here are important, but are not visible at this resolution. For example, figures to-

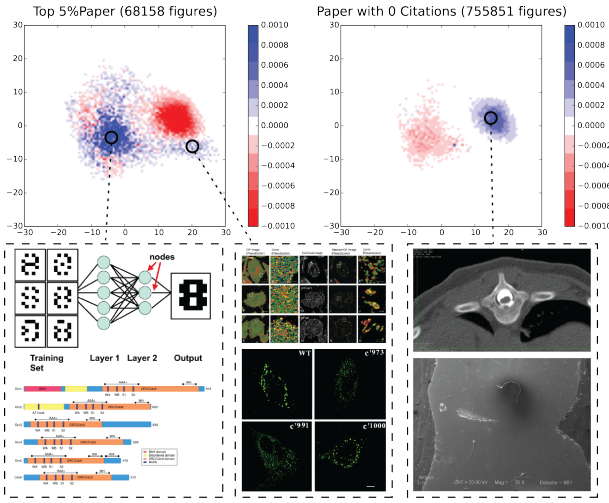

Figure 3: Deviations from the average distribution of figures from high-impact papers (top 5% as ranked by Eigenfactor) and low-impact papers (those that have attracted zero citations). Low-impact papers are characterized by photographic figures rather than data-oriented visualizations or scientific illustrations.

ward the upper right tend to be more heavily processed photos, including microarray images and fluorescence plots. The middle of the cluster, which we sample in Figure 1, tends to contain more natural photographs.

Although this initial figure provides a rough sense of the features extracted by the neural network, it does not deliver much insight. In the remainder of this paper, we will consider how various subsets of figures (those from high impact papers, those from recent years, those from specific journals) deviate from this global pattern, and discuss how those deviations might be interpreted. Further, we show how this kind of analysis of the space of visualizations can be used to bootstrap information extraction projects for specific figure types.

All images were rescaled to 224x224 (keeping aspect ratio and padding smaller images with empty space) and fed through the first layers of a pre-trained ResNet model [12]. ResNet model was pre-trained on ImageNet dataset.

4. TRENDS OVER TIME

In Figure 2, we show the deviation from the year-weighted average for two particular years in the corpus: 2006 and 2012. Blue regions represent a positive deviation from the average and red regions represent a negative deviation. We use the year-weighted average to account for the differences in publication volume year-to-year; without this correction, later years would appear to show a relative increase in all regions of the map. We see a relative increase in rich, complex diagrams, scientific illustrations, and dense imagery, and a relative decrease in quantitative plots.

We attribute this increase to software improvements that have made it easier to create high-quality diagrams and illustrations.

It is tempting to assume that this shift indicates that diagrams, illustrations, and imagery are associated with improved communicability or scientific impact; after all, why would their use be increasing if they were not effective? However, when we construct residual plots based on impact, we see a different pattern, which we now describe.

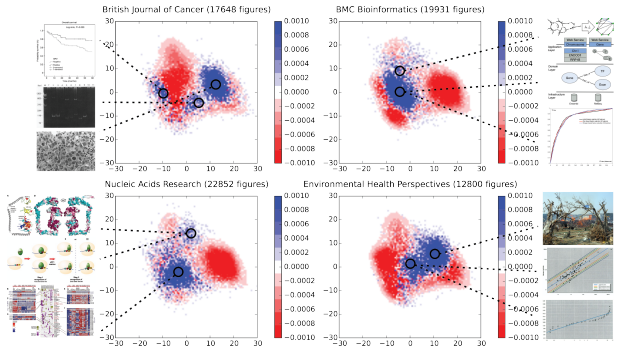

Figure 4: Deviations from the average distribution of the figures from four different journals: British Journal of Cancer, BMC Bioinformatics, Nucleic Acids Research, and Environmental health perspectives.

5. RELATIONSHIP TO IMPACT

In Figure 3 we consider how the figures from the top 5% of all papers ranked by Eigenfactor score [3] differ from the global baseline. We contrast this set of high-impact papers with the set of all papers that received zero citations.

The differences are greater for the high-impact papers because there are significantly more zero-citation papers. That is, most papers have zero citations, so we would expect the integral of the residuals to be smaller than that of high impact papers; this appears to be the case.

Multiple patterns are clear: Among high-impact papers, there are significantly more plots and diagrams (left-hand blue region) and significantly fewer photographs. High-impact papers exhibit a slight increase in a region at the lower right. This region is visually similar to photographs, but contain heavily processed images, including those from fluorescence experiments and microarray experiments.

Our interpretation of these results is that high-impact papers are associated with empirical results and, potentially, high-quality presentation of these results. This interpretation is primarily useful as a sanity check on our methods and visualization. A more detailed analysis of this space to understand the additional structure in Figure 3, which we will pursue as part of future work, will help us understand how presentation of data related to effective scientific communication. We hope to use these methods to make actionable recommendations for researchers in presenting their results both to interdisciplinary audiences and intradisciplinary audiences.

6. VISUAL SIGNATURES BY JOURNAL

In Figure 4, we show the residual plots for four specific journals to understand how these methods can help expose the *visual signature* of a particular publication venue. We posit that these signatures can help expose how different disciplines understand and communicate complex ideas. More practically, we envision tools that can help authors pre-

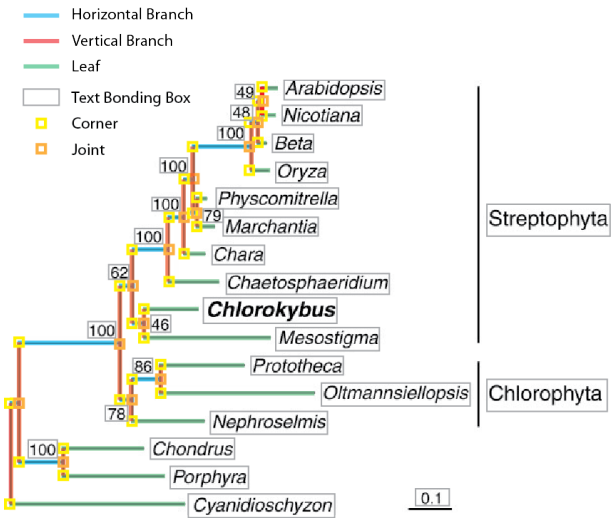

Figure 5: Figures illustrating phylogenetic trees carry information otherwise unavailable in the text of the paper. We can learn to identify phylogenetic trees using the unsupervised methods we describe without training data. Once a sample is available, we can develop algorithms to extract the structure and labels to construct a longitudinal database of phylogenetic databases.

pare visualizations that will conform to particular styles and therefore maximize their communicability to reviewers and readers in particular communities. An important overall goal is to enhance scientific communication between disciplines and with the general public, and we see the visual literature as an important mechanism mediating such communication.

At the upper left, the British Journal of Cancer is characterized by photographic imagery (microscopy in particular), and a particular type of data visualization that includes survival curves. In contrast, BMC Bioinformatics (upper right), is characterized by data and diagrams, suggesting the importance of computational abstractions in the research: systems architecture, illustrations of the operations of algorithms, etc.

At the lower left, Nucleic Acids Research emphasizes richer data-intensive visualizations, including heatmaps for gene expression data. But also, rich scientific illustrations of proteins and genetic structures are common.

Finally, at the lower right, the journal Environmental Health Perspectives includes significant photographic data, and may be associated with a particular "style" of plot with a darker background.

The distinctions between these journals are remarkably clear, and the patterns exhibited are consistent with our intuition about the emphasis of each venue. This consistency suggests that this method could be used to relate visual patterns to styles of scientific communication across fields and their relative efficacy.

7. VISUAL INFORMATION EXTRACTION

Using the unsupervised methods we describe, samples of specific categories of figures can be readily identified for further analysis. In particular, certain types of images contain information that is otherwise unavailable in the text of the paper. For example, metabolic pathway diagrams summarize hundreds of experiments visually to show the relationships between molecular and cellular processes.

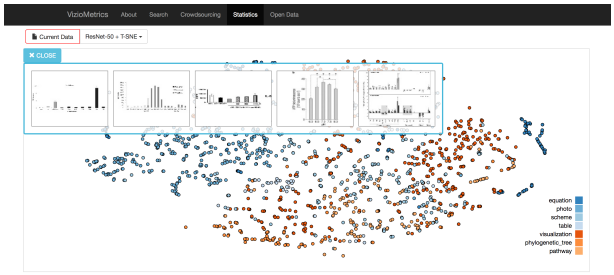

Figure 6: Screenshot of an interactive browser for a sample of the figure dataset. Each point is a figure; the axes are defined by various clustering and dimension-reduction methods.

As part of the Viziometrics project, we are extracting information from phylogenetic tree diagrams (Figure 5). These diagrams encode the results of computational and wet-lab experiments to organize the evolutionary relationships between different species. By identifying the set of all such figures in the literature, we aim to construct a longitudinal database of phylogenetic information, evaluate it against public phylogeny resources, and use the results to make inferences about coverage and consistency in the literature. For example, significant disagreement or gaps in certain areas of the tree of life could motivate new research.

In Figure 5, the left-hand side shows an example of the kind of figure our algorithms are able to parse. On the right, we summarize the major steps in the algorithm. The algorithms use machine vision techniques to identify corners and joints to form branches, then assemble branches into collections, then use tracing techniques to connect collections into subtrees. The subtrees are then connected to leaf labels extracted using OCR techniques to form complete phylogenetic trees.

This project is just one example of the kind of research we envision supporting with the visual map of the literature.

8. VIZIOMETRICS ONLINE

We have developed a set of online tools based on these methods for working with viziometrics data.

Figure 6 is a screenshot of an interactive browser for a sample of the dataset of figures. This interface is designed to help quickly understand the behavior of the clustering and dimension reduction methods: When structure is apparent, what does it mean qualitatively? To facilitate this kind of analysis, the user can hover over an individual point to view the figure it represents. A lasso interaction over a set of points displays all of the selected figures as thumbnails. In Figure 6, the user has selected a set of data visualizations forming a tight cluster; the visual similarity in this tight cluster is clear.

The colors of the points represent the coarse-grained labels assigned by the supervised classifier presented in prior work. In this context, these labels help sanity check the

clusters, guide user interaction, and in some cases uncover misclassifications in the supervised framework.

9. CONCLUSIONS

We developed a pipeline for mapping the visual literature using a pre-trained neural network and PCA, and used this pipeline to present preliminary results that show how patterns in figure use have changed over time, relate to impact, and vary by field.

We consider these methods to be a baseline for new research in how the use of visualization influences scientific communication within and across fields.

These methods also enable targeted information extraction projects on specific figure types; we are currently exploring algorithms for extracting information from phylogenetic trees.

10. REFERENCES

[1] A. Ahmed, A. Arnold, L. P. Coelho, J. Kangas, A. S. Sheikh, E. Xing, W. Cohen, and R. F. Murphy. Structured literature image finder: Parsing text and figures in biomedical literature. *Journal of Web Semantics*, 8:151–154, 2010.

[2] A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48. ACM, 2009.

[3] C. T. Bergstrom, J. D. West, and M. A. Wiseman. The eigenfactorâĎć metrics. *Journal of Neuroscience*, 28(45):11433–11434, 2008.

[4] G. Cabanac, G. Hubert, and J. Hartley. Solo versus collaborative writing: Discrepancies in the use of tables and graphs in academic articles. *Journal of the Association for Information Science and Technology*, 65(4):812–820, 2014.

[5] Z. Chen, M. Cafarella, and E. Adar. Diagramflyer: A search engine for data-driven diagrams. In *Proceedings of the 24th International Conference on World Wide Web Companion*, WWW '15 Companion, pages 183–186, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[6] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[7] S. R. Choudhury, P. Mitra, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles. Figure metadata extraction from digital documents. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 135–139. IEEE, 2013.

[8] S. R. Choudhury, S. Wang, P. Mitra, and C. L. Giles. Automated Data Extraction from Scholarly Line Graphs. 2013.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[10] T. W. Fawcett and A. D. Higginson. Heavy use of equations impedes communication among biologists. *Proceedings of the National Academy of Sciences*, 109(29):11735–11739, 2012.

[11] J. Hartley and G. Cabanac. Do men and women differ in their use of tables and graphs in academic publications? *Scientometrics*, 98(2):1161–1172, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *ArXiv e-prints*, Dec. 2015.

[13] P. Hegarty and Z. Walton. The consequences of predicting scientific impact in psychology using journal impact factors. *Perspectives on Psychological Science*, 7(1):72–78, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[15] P. Lee, J. West, and B. Howe. Viziometrics: Analyzing visual information in the scientific literature. *arXiv preprint:1605.04951*, 2016.

[16] R. F. Murphy, M. Velliste, J. Yao, and G. Porreca. Searching online journals for fluorescence microscope images depicting protein subcellular location patterns. In *Bioinformatics and Bioengineering Conference, 2001. Proceedings of the IEEE 2nd International Symposium on*, pages 119–128. IEEE, 2001.

[17] R. A. Poldrack and T. Yarkoni. From brain maps to cognitive ontologies: Informatics and the search for mental structure. *Annual review of psychology*, 67:587–612, 2016.

[18] S. Ray Choudhury and C. L. Giles. An architecture for information extraction from figures in digital libraries. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 667–672. International World Wide Web Conferences Steering Committee, 2015.

[19] M. Tartanus, A. Wnuk, M. Kozak, and J. Hartley. Graphs and prestige in agricultural journals. *Journal of the American Society for Information Science and Technology*, 64(9):1946–1950, 2013.

[20] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.