A case study of viziometrics: What's the role of western blots in Alzheimer's Disease literature?

Satoshi Tsutsui[1], Zheng Gao[1], Zheng Wang[2], Guilin Meng[3], Ying Ding[1,4]

[1] School of Informatics, Computing, and Engineering, Indiana University, USA
[2]Department of Information Management, Nanjing University of Science and Technology, China
[3]Department of Neurology, Tenth Peoples Hospital,Tongji University, Shanghai, China
[4]School of Information Management, Wuhan University, Wuhan, China
{stsutsui,gao27,dingying}@ndiana.edu, wangyz@njust.edu.cn

Abstract. The visual information in scientific could play an important role, but few bibliometric studies investigate it. In order to emphasize the importance of the visual aspect of scholarly communication, a new field called *viziometrics* is recently proposed. This paper presents an ongoing project for a case study of viziometrics where we focus on western blots within Alzheimer's Disease (AD) literature. We first develop a computer vision method to detect western blots from the images of figures. Then we extract thousands of western blots from AD papers and show a preliminary analysis.

Keywords: viziometrics, bibliometrics, computer vision

1 Introduction

Scientific papers use figures in order to efficiently demonstrate the findings. The visual information is much easier for humans to understand than textual information and thus could play a more important role than texts. The visual aspects of scholarly articles are relatively less explored than texts and citations, but recent studies proposed to explore "the organization and presentation of visual information in the scientific literature" [8, 5]. They call it *viziometrics*, which is to emphasize the study of visual information in scholarly articles with the similar goal of bibliometrics. These studies investigate general biomedical fields in an aggregated way and only consider general types of figures such as tables, diagrams, photos, and plots. In contrast, the purpose of this study is to investigate how a scientific community uses a specific type of figure. We believe that narrowing down the focus will result in an interesting discovery on the role of visual information, providing a novel case study on viziometrics. For example, papers in information science could represent results of topic modeling algorithm with intuitive diagrams wheres papers in computer science might prefer to use tables with quantitative metrics when presenting the results of the same algorithm, which would be a specific instance of (visual) "cultural norms for publication" [8].

In this study, we analyze western blots from Alzheimers Disease (AD) papers. AD is a neurodegenerative disorder, which is specific enough to fit the purpose of our study, and also has been a target of bibliometric studies due to its societal impact and the large volume of the related papers [1, 7, 10]. In fact, just searching *Alzheimer's Disease* in PubMed gives more than 10,000 papers. However, previous work only uses citations and texts, and have not explored visual information. We particularly focus on the western blot (an example is shown in Figure 1), which is a widely used for analyzing the protein expression in the biomedical domain. In fact, our study shows that 37% of the AD papers contain western blots. This fact indicates that the western blot fits our purpose because it is a widely used domain specific figure that can be used in various contexts, whose different usages could constitute latent visual cultures that is not discovered yet.

This study requires a computer vision algorithm to detect western blots from large numbers of figures extracted from AD papers. The state-of-the-art computer vision is supported by deep learning [6], but it requires thousands or millions of annotated images, which is a challenge for us. We address the challenge using a machine learning framework called transfer learning [12], and are able to build a high-performance western blot recognition algorithm while we prepare only a thousand training images. We released a easy-to-use python code[1] to recognize western blots in order to facilitate more research in viziometrics.

This paper describes an ongoing project of a viziometrics case study: revealing the role of western blots within AD papers. We first explain our approach for developing the western blot recognition algorithm using computer vision techniques (§2). Then we report how we collect western blots from AD papers (§3). We finally discuss preliminary results and the future directions (§4).

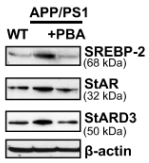

Fig. 1. An example of Western Blots

2 Western Blot Recognition Method

We adopt a machine learning approach and train a deep convolutional neural network called ResNet50 [4] to classify western blots or not given an image of a figure. It requires training images (i.e., figures labeled western blots or not), which we collect from a publicly available dataset of biomedical figures [3]. We manually investigate the images labeled as *Chromatography* and obtain 544 western blots. We also obtain random 544 figures that are not western blot, and use them as negative training data. From the total 1,088 images, we train

[1] The code is available from: `https://github.com/apple2373/western_recognize`

ResNet50 using 762 randomly sampled images (i.e., training set), and evaluate the performance using the remaining 326 images (i.e., test set). Each set has the same number of western blots and figures of non western blot. A challenge is, the number of images is much smaller than the number that is typically required for training neural networks [6]. We overcome the challenge by a transfer learning approach introduced in [12] using ImageNet [9]. That is, we first train general classification model using 1.2 million images, and then reuse the learned parameters as the initial parameters for training western blot classifier. This gives the classification accuracy of 97.1%. We implement the method using python and open sourced the resulted classifier[1], hoping to foster more research in this domain.

3 Western Blot Extraction from AD Literature

We first obtain PDFs of AD papers, and extract figures from them. We retrieve 110,321 PDFs of the papers published in 2007-2016 from the PubMed Central using the search words of "Alzheimer's Disease". From these PDFs, we extract 415,683 figures using PDFFigures [2]. Some of the figures are compound figures (i.e., a figure comprised of multiple figures), so we decomposed into individual figures using a compound figure separation tool [11], and obtain 1,208,663 figures. From them, 168,490 western blots are identified using the computer vision method described in § 2.

4 Preliminary Results and Discussion

Our final goal is to reveal the role of western blots within AD papers. We first discover that 40,904 papers which are 37% of the AD-related publications we collected, evidencing that western blots are one of the fundamental figures used in AD research papers. Then, in this preliminary study, we focus on a simple question: When western blots are used in AD papers? To answer the question, we use journals where papers are published, and Medical Subject Headings (MeSH) terms, which are the keywords given to papers.

We select major journals by only using ones that have at least 1,000 papers and compute the proportion of papers that have western blots for each journal. The results are are *Journal of Biological Chemistry*: 83%, *Journal of Neuroscience*: 72%, *Scientific Reports*: 58% , *PLOS ONE*: 53%, *Proceedings of the National Academy of Sciences (PNAS)*: 49%, and *NeuroImage*: 30%. This suggests that the western blots are often used in biological chemistry and neuroscience but not so often in neuroimaging research. We also see that almost 50% of AD papers published in general journals such as PLOS ONE have western blots. This is another evidence that western blot is an important figure in AD research.

We first tried the similar approach for MeSH terms but it did not work well because top frequent terms are too general (e.g, *Humans*) and not so informative. The similar issue is also reported in the previous study where manual selection

of MeSH terms is performed [7]. In this preliminary work, we use correlation coefficients between the appearance of western blot and the appearance of each MeSH term, based on the hypothesis that general terms tend to appear in papers with a variety of topics, making the coefficients nearly zero. The top five positively correlating MeSH terms are *Mice, Animals, Cells(Cultured), Mice(Inbred C57B)*, and *Rats*. This indicates that western blots might often be used in the research involving experiments with mice/rats.

We, of course, do not claim that the results with MeSH terms and simple correlations are enough. We would like to identify more precise contexts where western blots are used. They could be the sentences and paragraphs where western blots are pointed and presented with their interpretations. Investigating these contexts would distinguish the multiple contexts that depend on the arguments where the authors would like to present. Moreover, we would like to answer more interesting questions such as; What kind of figures are frequently used with western blots? (i.e., co-figure analysis).

References

1. H. Chen, Y. Wan, S. Jiang, and Y. Cheng. Alzheimer's disease research in the future: bibliometric analysis of cholinesterase inhibitors from 1993 to 2012. *Scientometrics*, 98(3):1865–1877, 2014.
2. C. Clark and S. Divvala. Pdffigures 2.0: Mining figures from research papers. In *The Joint Conference on Digital Libraries (JCDL)*, 2016.
3. A. García Seco de Herrera, R. Schaer, S. Bromuri, and H. Müller. Overview of the ImageCLEF 2016 medical task. In *Working Notes of CLEF*, 2016.
4. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
5. B. Howe, P. Lee, M. Grechkin, T. S. Yang, and J. West. Deep mapping of the visual literature. In *International Conference on World Wide Web (WWW) BigScholar Workshop*, 2017.
6. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
7. D. Lee, W. C. Kim, A. Charidimou, and M. Song. A bird's-eye view of alzheimer's disease research: reflecting different perspectives of indexers, authors, or citers in mapping the field. *Journal of Alzheimer's Disease*, 45(4):1207–1222, 2015.
8. P. Lee, J. West, and B. Howe. Viziometrics: Analyzing visual patterns in the scientific literature. *IEEE Transactions on Big Data*, 2017.
9. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
10. M. Song, G. E. Heo, and D. Lee. Identifying the landscape of alzheimer‘s disease research with network and content analysis. *Scientometrics*, 102(1):905–927, 2015.
11. S. Tsutsui and D. Crandall. A Data Driven Approach for Compound Figure Separation Using Convolutional Neural Networks. In *The IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.
12. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NIPS)*, 2014.