Analysis of the co-authorship network of Filipino researchers in deep learning

Angelyn R. Lao*[1,2], Unisse C. Chua[2,3], and Briane Paul V. Samson[2,3,4]

[1]Mathematics and Statistics Department, De La Salle University, 0922 Manila, Philippines
[2]Center for Complexity and Emerging Technologies, De La Salle University, 0922 Manila, Philippines
[3]Software Technology Department, De La Salle University, 0922 Manila, Philippines
[4]Future University Hakodate, Hakodate, Japan

Deep learning is an emerging field that attracts attention from various researchers to use in analyzing and understanding increasingly complex systems and phenomena. Crucial to the advancement of deep learning research in the Philippines, institutions need to understand its direction and reach out and work with other leading researchers who have knowledge, insight, and resources to contribute to the challenges in deep learning research. Due to its wide application, we analyzed the collaboration dynamics and patterns of Filipino researchers working in this field through their co-authorship network. Our study found that while there is a steady increase in research productivity on deep learning, most of the publications and collaborations are concentrated on a handful of institutions like De La Salle University and University of the Philippines Diliman. Despite their current control over the direction of the local research community, these two top institutions have yet to form strong collaborations. We also found that Filipino researchers are mostly doing applied deep learning projects and rely heavily on undergraduate students to maintain productivity.

KEYWORDS

co-authorship, network analysis, research productivity, Filipino, deep learning, Philippines

*Corresponding author
Email Address: angelyn.lao@dlsu.edu.ph
Date received: April 1, 2020
Date revised: August 14, 2020
Date accepted: August 25, 2020

INTRODUCTION

Deep learning (DL), inspired by the structure and function of artificial neural networks (ANN), is a relatively new area of machine learning (ML). DL's algorithms are inspired from the human brain wherein the machines gain intelligence without explicit programming (Moolayil 2019). Brownlee (2019) described it as a technique that learns the best way to represent the problem while learning how to solve the problem, known as representation learning. This is what differentiates DL from classical ML models. DL models are said to be so powerful that they spur researchers to further develop, evaluate, and apply them. The caveat of DL models is to ensure the robustness of the results and ensure that the solutions are suitable for the problems.

Over recent years, there has been increasing interest among researchers in the notion of research collaboration (Katz and Martin 1997). It is widely assumed that collaboration in research is a good thing and that it should be encouraged. Crucial to the development of DL studies in the Philippines is for the national leading institutions to identify the directions of future DL research and to reach out and work with institutions that certainly need knowledge, insight, and resources to contribute to the challenges in DL studies. According to Adams and Loach (2015): "excellence talking to excellence is a powerful force, but the net benefit may come from a slightly wider discourse."

In this paper, we use traditional indicators such as number and citation count of the published scientific articles to get an overview of the current development and directions of the Filipino researchers in DL. We would like to understand the collaboration dynamics or patterns of Filipino researchers in DL through their co-authorship networks. Extracting these patterns among successful researchers might be beneficial to both researchers and decision makers who are seeking to improve the researchers' productivity (Al-Ayyoub et al. 2017; Adams and Loach 2015). We aim at finding answers to questions like: Which institutions are part of this growing network? How do the DL researchers interact with each other? Do they work individually or in groups? If it is the latter, do they collaborate with local or foreign researchers? Who they work with? Are there special collaboration patterns common between highly productive researchers? How does their research contribute to innovation?

MATERIALS AND METHODS

Collecting publication data

The publications of Filipino researchers working on deep learning were retrieved in two ways: through (1) Elsevier's Scopus Search API[1], and (2) direct solicitation from the authors. Elsevier provides access to abstracts and citation data of all Scopus-indexed scholarly articles through public APIs. The Scopus Search API provides the same functionality as the online article search. To collect data on the publications, we used the search query *'ALL ("deep learning") AND AFFILCOUNTRY(Philippines)'*. This query retrieves all articles that have at least one author affiliated with the Philippines and has the exact phrase "deep learning" in any of the fields, including the abstract, title, and keywords. We collected articles published until January 2020. This resulted in 193 articles, with the oldest one published in 2006. However, not all of them were about deep learning in computer science context. Ninety-eight (98) of them were removed after manual inspection because they were about deep learning in the context of education. This leaves out 95 articles for analysis. Additionally, from the direct solicitation, there were additional 13 articles from Scopus that were not retrieved by our search term and another 21 articles that were not indexed on Scopus.

We were able to retrieve the complete metadata[2] of the articles provided by Scopus through our institution's subscription. Since the search only returns affiliation IDs as part of the result, we used the Affiliation Retrieval API[3] to include the actual names of the authors' affiliation. In the case of an author having multiple affiliations, we only considered the first affiliation provided by the API, which is usually the most recent. By doing this, we might be missing information on some authors who are Filipino but published an article while completing their graduate studies abroad. Another limitation in our data is that the affiliation considered is the affiliation associated with the published article.

After merging the data collected, we manually assigned the Scopus author ID for the articles that were directly solicited. For co-authors that were not on Scopus, we provided them with a unique random ID.

In order to generate a network, we created two comma-separated value (CSV) files. One CSV contains information on the unique authors which include a unique identifier, their full name, and affiliation. The other file primarily contains all author-author pairs per unique publication. An article that has 5 authors will have 10 author-author pairs. Each row in this file contains the paper ID and the unique IDs of two authors.

Building the co-authorship network

The co-authorship network represents the relationships between Filipino researchers after they publish articles together. Each node represents a unique author with at least one published article on deep learning. Connecting two authors are edges that indicate the presence of at least one publication they have co-authored. These edges are undirected which indicates a symmetrical relationship. Also, each edge between authors has weights based on the total number of articles they have written together. The initial network from the co-authorship data yielded 295 nodes and 640 edges. Then, we removed the nodes of researchers who are affiliated with a non-Philippine institution. The final co-authorship network now contains 261 nodes and 529 edges. In our dataset, we have 10 articles with only a single author. These articles are included in the main analysis of papers published by Filipino researchers, however they are excluded in the co-authorship network. The authors are still included as nodes in the network since they have other published articles with other researchers.

Analyzing the network structure

First, we characterized the co-authorship network by calculating its centrality measures. In order to determine how the researcher actively collaborates with other researchers, we computed the total number of edges connected to each node or their ***degrees***. Because not all edges are the same, we also computed the sum of their weights or their ***weighted degrees***. These two centrality measures allowed us to determine hubs within the network or the central researchers.

We also computed the ***shortest paths*** between any two nodes. Then, we identified the ***network diameter*** or the longest among the computed shortest paths. This gives the linear size of the network.

Identifying communities

Real-world networks, despite varying degrees of connectivity, often form well-connected groups of nodes called communities. To identify the well-connected research communities within the co-authorship network, we used the ***Leiden community detection algorithm*** (Traag et al. 2019) instead of the commonly used Louvain method. The Leiden community detection algorithm is composed of three phases, which are iterated over until no more improvements can be made. The first phase is the local movement of nodes. It starts by partitioning the whole network into singletons (a list with only one element) and then moving individual nodes from one community to another to find a partition. The second phase is the refinement of the partition. This may involve making subcommunities from non-refined partitions. The last phase is the aggregation of the network based on the refined partitions. The aggregated network contains nodes that represent the refined partitions from the previous phase. Then, initial partitions of the aggregated network are based on the non-refined partitions.

Measuring research productivity

In order to determine the distribution of publications and measure the research productivity of a research group, the ***average productivity research index*** (APRI) is measured by computing the ratio of publications to authors (Abramo and D'Angelo 2014).

Uncovering topics of interest

Structural Topic Model (STM) is performed to discover the topics of interest of the Filipino DL researchers and estimate their relationship to document metadata (Roberts et al. 2014). STM is a statistical and machine-assisted approach to read and analyze text corpora. The foundation of STM is based on

[1] https://dev.elsevier.com/documentation/ScopusSearchAPI.wadl

[2] https://dev.elsevier.com/guides/ScopusSearchViews.htm

[3] https://dev.elsevier.com/documentation/AffiliationRetrievalAPI.wadl

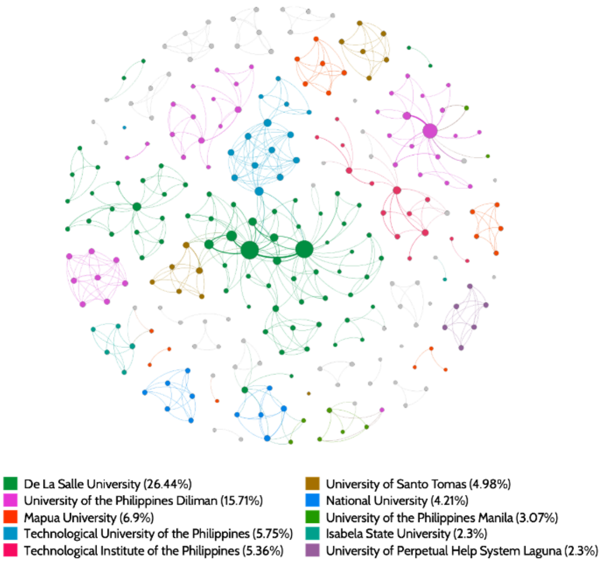

Figure 1: The whole network of Filipino researchers with publications on deep learning. Each node is a researcher with its size representing the number of connections they have with other nodes and its color denoting their current affiliation. Only the top 10 institutions with the most researchers in the network are encoded with specific colors. Researchers from other schools are colored gray.

probabilistic topic models, such as the Latent Dirichlet Allocation (LDA).

Here we briefly describe the generative process and estimation of STM. The generative model begins with document-topic and topic-word distributions generating documents that have metadata associated with them. In a mixture of words, a topic is defined based on the probability of a word belonging to a topic. In a single document, it can be composed of multiple topics. As such, the sum of the topic proportions across all topics for a document is one, and the sum word probabilities for a given topic is one (Roberts et al. 2014). Some features of STM include the computation of topical prevalence and topical content. Topical prevalence refers to "how much of a document is associated with a topic" and "topical content refers to the words used within a topic" (Roberts et al. 2014). For this study, we used the R package named *stm*.

RESULTS AND DISCUSSION

We assessed and analyzed the scientific productivity of Filipino authors and institutions in peer-reviewed journals. The publications were retrieved from Scopus using the queries described in Materials & Methodologies.

Key Institutions

Identifying the key institutions and researchers is important, as they are responsible for keeping several other institutions in the loop and should, therefore, be considered as fundamental partners for training, capacity building and institutional strengthening (Morel et al. 2019). Figure 1 shows the Filipino DL co-authorship network which is made up of researchers from 39 Philippine higher-education institutions and 1 national institute. The largest group working on deep learning are from De La Salle University with 71 researchers (26.4%). They are followed by the University of the Philippines Diliman with 43 researchers (15.7%) and Mapua University with 18 (6.9%). On the other hand, there are 19 institutions with lone researchers working and or collaborating on deep learning projects.

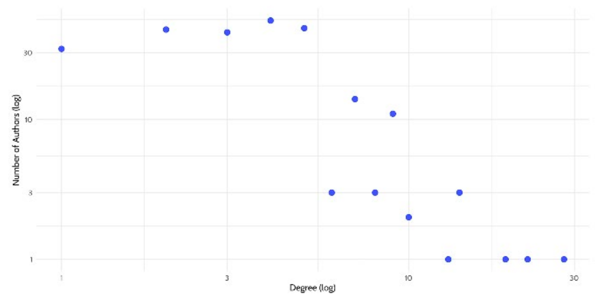

Figure 2: A log-log plot of the degree distribution of nodes in the network. It somehow follows a typical distribution of a real world network where there are many nodes with low degrees and only few with very high degrees. Nodes with zero degrees were removed in the plot.

Hubs of Collaboration and Network Characteristics

The network has an average degree of 4.05 which means Filipino DL researchers have collaborations with 4 other people on average. If we consider how often pairs of researchers work with each other, the network's average weighted degree of 4.57 suggests that for the majority, on the average, they co-author only once. Figure 2 shows the distribution of degree values which follows a typical distribution for a real-world network. The log-log plot shows that the majority of researchers have very low degrees while only a few researchers have many strong connections. Acting as hubs of collaboration in the network, Melvin Cabatuan ($k = 28$) and Elmer Dadios ($k = 22$) of DLSU, and Prospero Naval ($k = 19$) of UPD have the most connections. Researchers with high degrees are mostly from DLSU, UPD and Technological University of the Philippines (TUP). On the other hand, 32 researchers collaborate with only one other person ($k = 1$) and seven are working alone ($k = 0$).

When we look closer into each institution (Figure 3), the weighted degree distributions of De La Salle University (DLSU) and University of the Philippines Diliman (UPD) mimic that of the whole network. Four out of the top five researchers with high weighted degrees are from DLSU which suggests a highly collaborative effort with multiple groups over time. This time around, Elmer Dadios recorded the highest weighted degree ($k_{weighted} = 44$) while Melvin Cabatuan had one less ($k_{weighted} = 43$). This suggests that some researchers have built stronger and more frequent collaborations with other researchers despite having fewer connections.

Between any two nodes in the network, the average path length was 2.37. Considering all shortest paths, the network has a diameter of 5. Looking at the overall network structure, it is indicative of the common practice of working with mostly undergraduate students who only get involved in doing one-time thesis projects and then join the industry after graduation. Although a number of them pursue graduate studies and or research careers, this practice allows few opportunities to extend their collaboration network.

Research Communities

The author nodes in the network are critical in connecting modular subnetworks of a network. Using the Leiden algorithm, we have shown clusters/subnetworks of Fipinino researchers who are well connected in the co-authorship network. Overall, we have identified 43 clusters. From Figures 4 and 5, the largest is Cluster 1, composed mostly of DLSU researchers with Cabatuan and Dadios as hubs. Everyone in this cluster has an engineering background. From UPD, the largest subnetwork (Cluster 2) is led by Naval working with other computer scientists. The other three big subnetworks are research groups led by Medina in TUP (Cluster 3), Cordel in DLSU (Cluster 4) and Atienza and Tiglao in UPD (Cluster 5). It can also be observed that Cluster 1 has the most diversity, with three main institutions working together namely DLSU, TUP and UST. The rest are dominated by one institution.

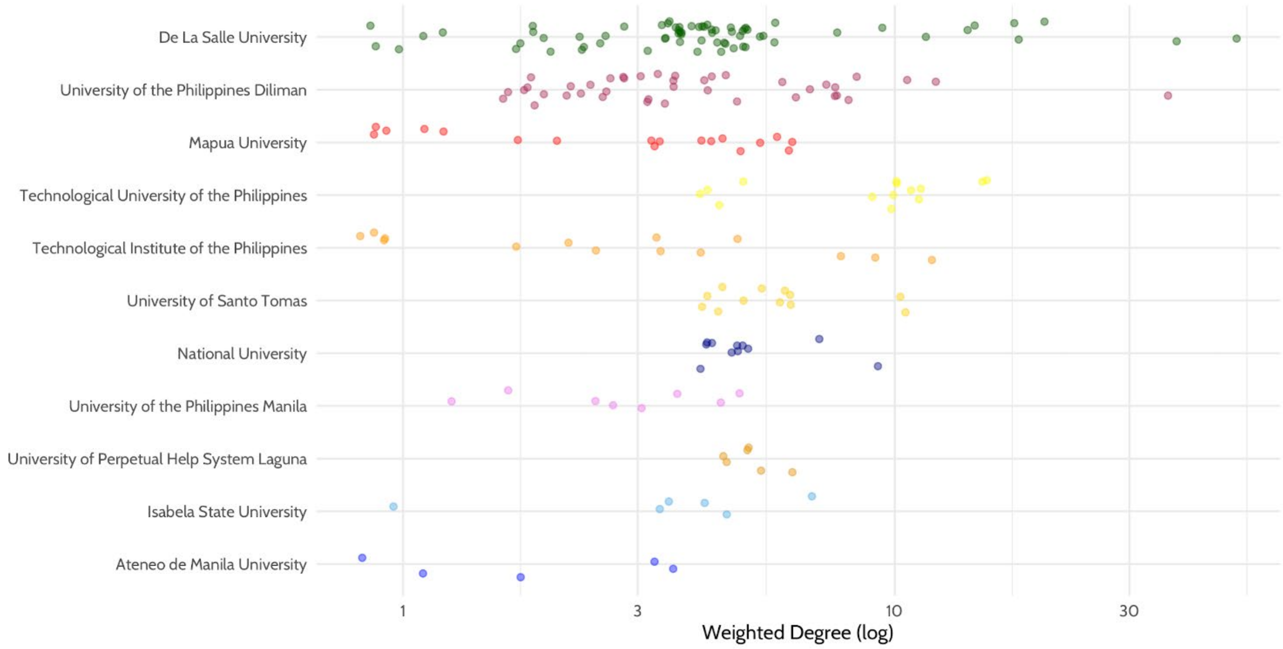

Figure 3: The weighted degree of researchers from the top eleven institutions with the most number of uniquely affiliated authors. Each dot represents a researcher and colors denote their affiliation. Larger weighted degree means that the majority of the works done by the researcher are done in collaborative effort.

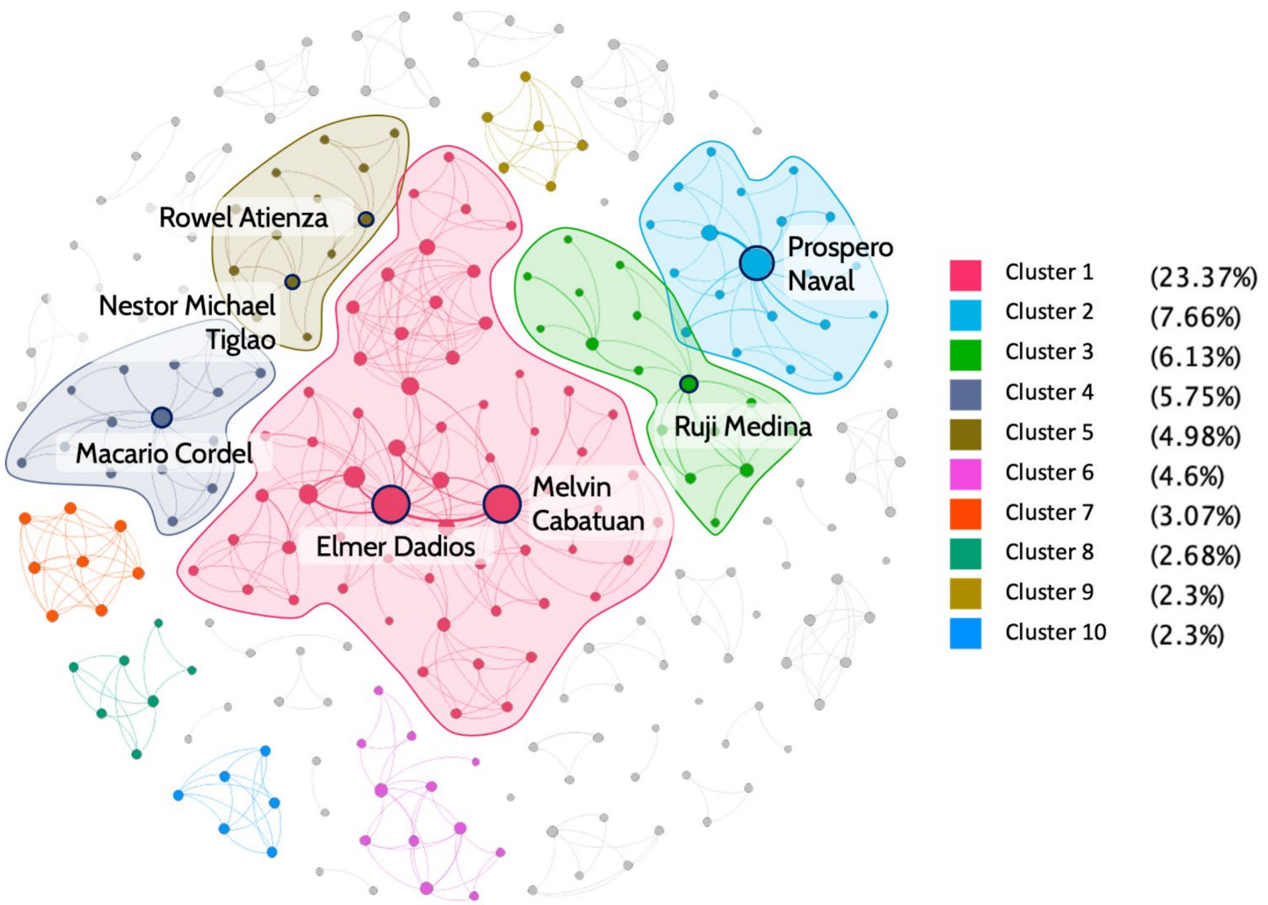

Figure 4: The network now shows the clusters or subnetworks of Filipino researchers who work closely together. Only the top ten clusters based on the total number of node members are shown. The largest cluster, Cluster 1, is in the middle, composed of mostly DLSU researchers.

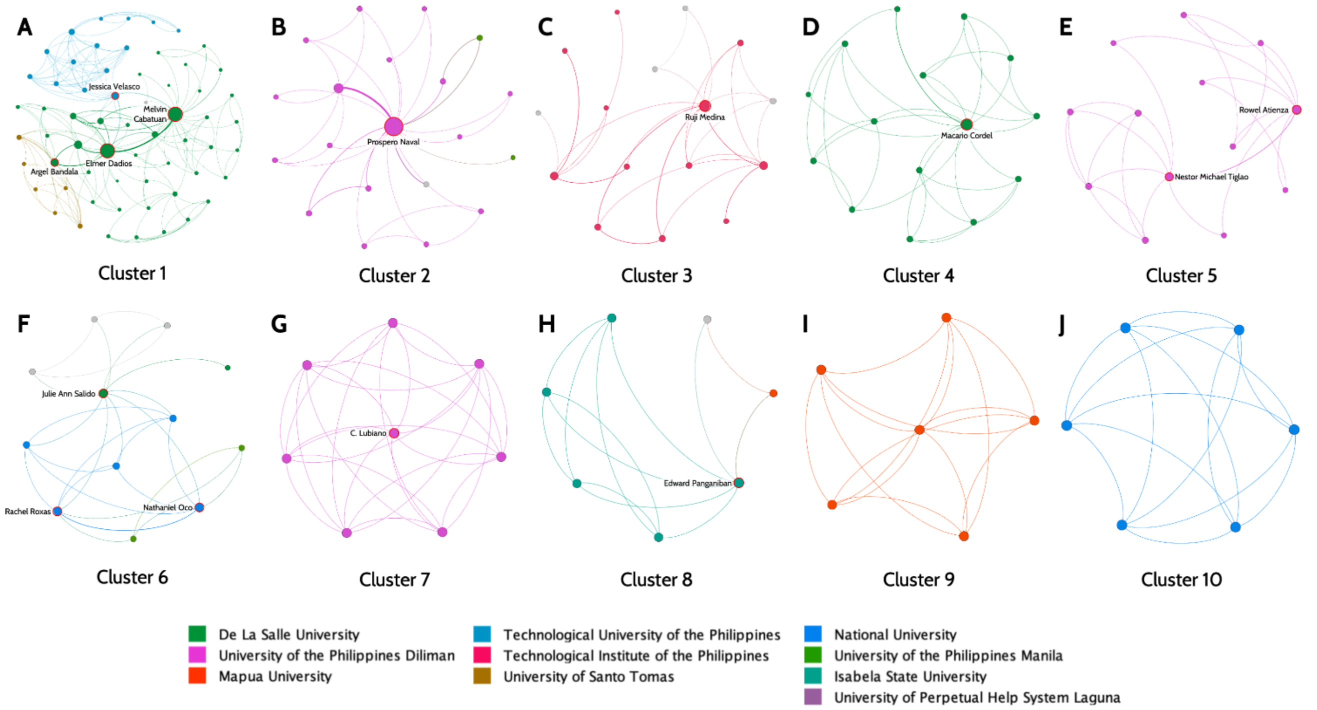

Figure 5: An isolated view of the top ten research clusters/subnetworks and their structure. The size of the nodes corresponds to their weighted degrees and are colored based on their affiliations. We labelled the nodes with the highest weighted degree for each cluster.

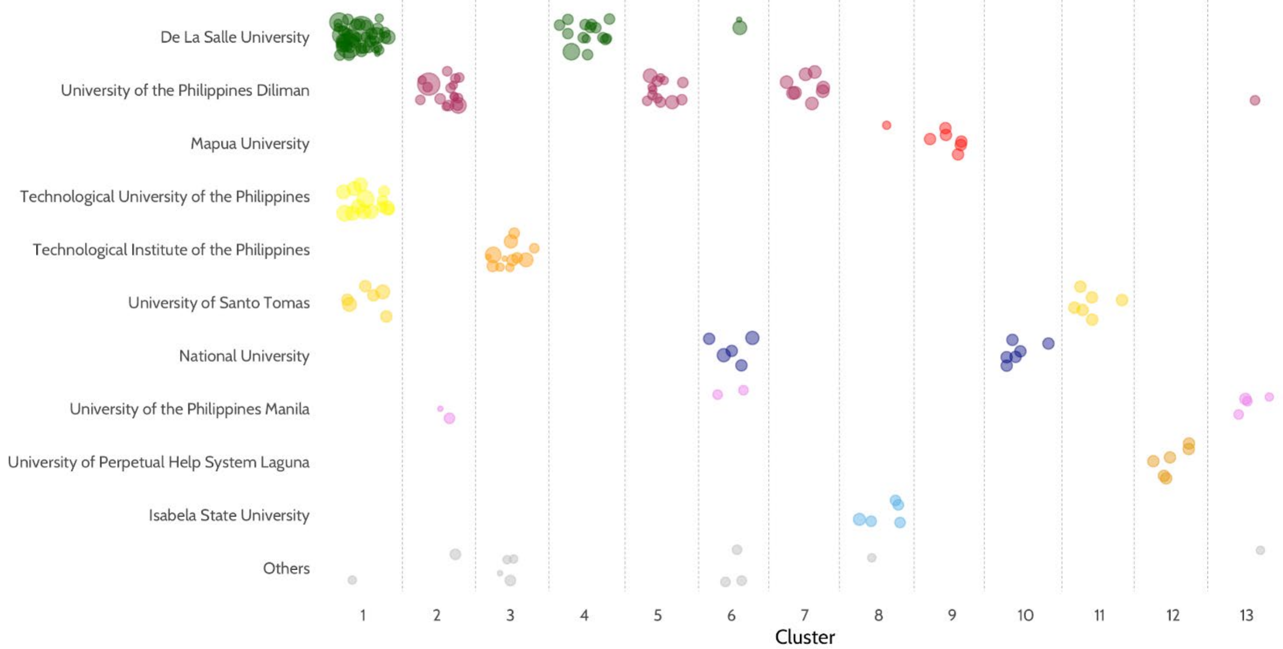

Figure 6: The distribution of authors from the top ten institutions to the top thirteen research clusters/subnetworks. The clusters are ordered from the research cluster with the most number of nodes (Cluster 1) to the least (Cluster 9 to 13). Clusters 9 to 13 have the same number of nodes (N = 6) so we included them all for comparison.

The pattern showing the distribution of authors from the top ten institutions to the top thirteen research clusters/subnetworks (shown in Figure 6) is also informative, likewise showing major variations. It is noticeable that researchers from DLSU, UPD and TUP are currently the major researchers of the three biggest research groups/clusters (Clusters 1 to 3). In terms of representation, UPD researchers are part of four top clusters -- the most among institutions. They are followed by DLSU and UPM with at least one researcher collaborating with others in three clusters. Interestingly, researchers from DLSU and UPD do not mix (collaborate) in any of the top research groups/clusters (listed here). Despite having many researchers in the network, Ateneo de Manila University (ADMU) researchers do not work with any of the top research groups working on deep learning.

How many of these collaborations are done locally (with all Filipino co-authors) and internationally (with foreign co-

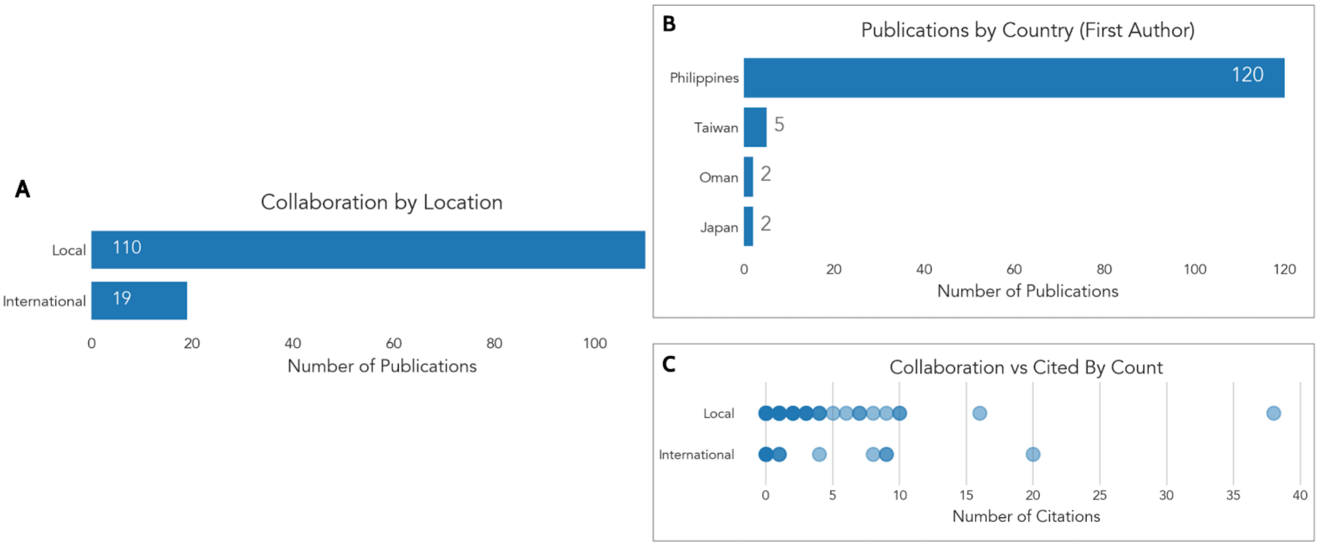

Figure 7: Publications with all Filipino co-authors and with foreign co-authors. (A) The chart shows the number of publications published locally (with all-Filipino authors) and internationally (with foreign-affiliated co-authors). (B) The chart shows the distributions of the affiliated country of the first author. (C) This chart shows the citation count of the publications published locally (with all-Filipino authors) and internationally (with foreign-affiliated co-authors).

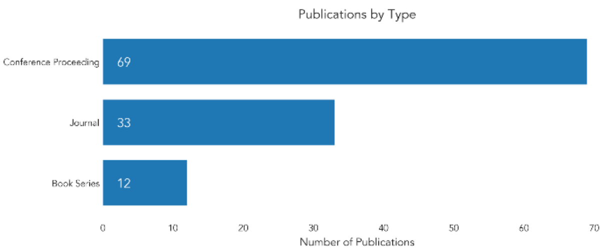

Figure 8: The chart shows the number of publications published in different types of publications.

authors)? Figure 7(a) shows that 110 of the papers published are purely made-in-the-Philippines with all-Filipino authors. Out of the 19 papers done with international collaborators/co-authors, 14 papers have Filipino as the first author (see Figure 7(b)). As shown in Figure 7(c), to have a high citation count, having international co-authors may not be relevant and may not be considered as an important factor.

Research Productivity

An observation regarding the publication trend in this field (illustrated in Figure 8) is that around 60% of the Filipino DL publications are submitted as conference proceedings. Only 29% is submitted for journal publications.

With the majority of the DL researchers coming from DLSU and UPD, it did not come as a surprise that they also topped the chart for being the most productive institutions in terms of DL publications (see Figure 9). What came as surprises are the following: (1) Isabela State University that is 10th in the "most number of researchers working on DL" came in 5th in the most productive institution; (2) Ateneo de Manila University and Aklan State University that is not even in the top ten "most number of researchers working on DL" came in 7th and 10th, respectively; and (3) Technological Institute of the Philippines who was the 5th "most number of researchers working on DL" surpassed Mapua University and landed as the 3rd most productive institution.

Based on the APRI (shown in Figure 9), UPM, TIP, and UPD have the highest index. The ratio of their publications to their researchers published in DL is relatively high as compared to other institutions. However, if it is based on the top 10 institutions, listed in Figure 2, with the most number of uniquely affiliated authors. The institutions with the top APRI are TIP, ISU, and UPD (shown in Figure 10). Consistently appearing in the charts are TIP and UPD. The median value of co-authors per paper is three (shown in Figure 13). Institutions with low APRI can also indicate a high collaboration rate between many unique authors, like the undergraduate students doing one-time thesis projects. It appears that a better indicator of good collaboration is when a three-author per paper median is obtained.

The top 10 researchers come from the top 3 leading institutions in DL research (see Figure 11), with Naval (from UPD) being the leading researcher in DL, joined by Cabatuan and Dadios from DLSU. Except for Atienza and Cordel, the other eight researchers have collaborative works done with one or more researchers in the chart. This suggests that there is an overlap in the publications being accounted for in this chart.

Figure 12 shows how many papers are published per year and how many citations it got since its publication (cut-off date is January 31, 2020, and the main source is SCOPUS). Since most of the papers are published very recently (60 publications in 2019), it is not surprising that around 90 publications still have low citation counts. For now, only the paper by Cordel in 2016 has a citation count of 38.

Topics of Interest

Applying structural topic modeling, we were able to identify 15 topic clusters (shown in Figure 14) with Topic 6 as the most prevalent across the published DL papers (shown in Figure 15). This topic is all about the use of "neural networks" in various application domains. This is followed by Topic 13 that talks about the use of "convolutional neural networks". However, there might be overlaps between Topics 6 and 13 because of the common/shared keywords "neural" and "network". The majority of the DL papers use it as a tool for machine learning projects. Very few pursue the fundamental and theoretical aspects of deep learning.

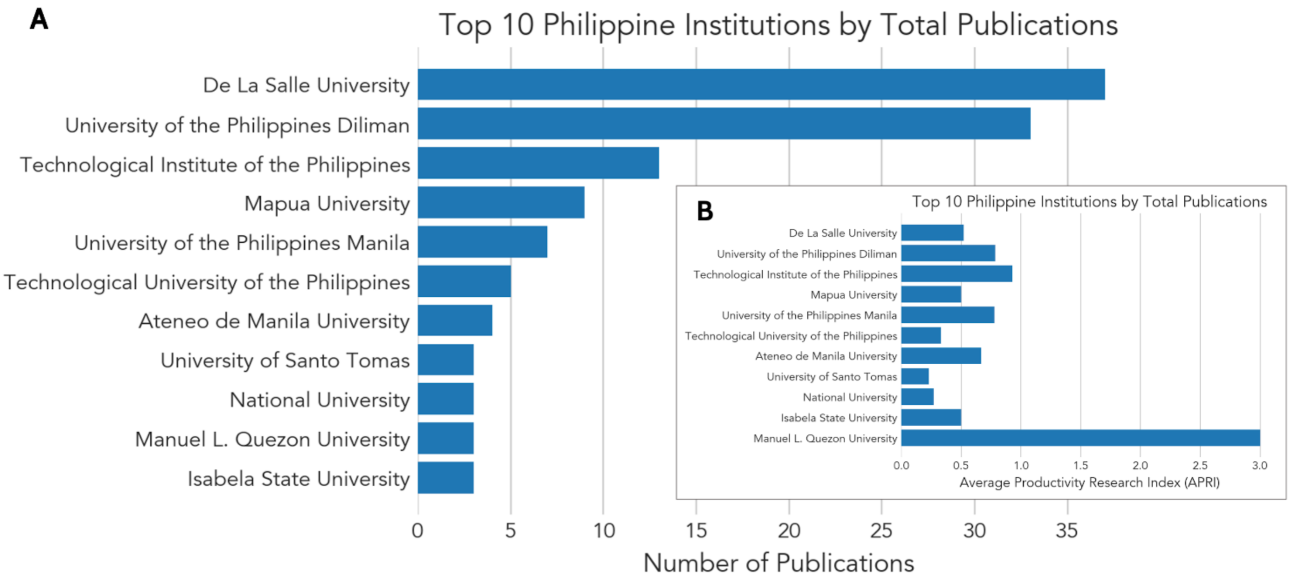

Figure 9: Top 10 Philippine institutions based on their publication count and corresponding average productivity research index (APRI).

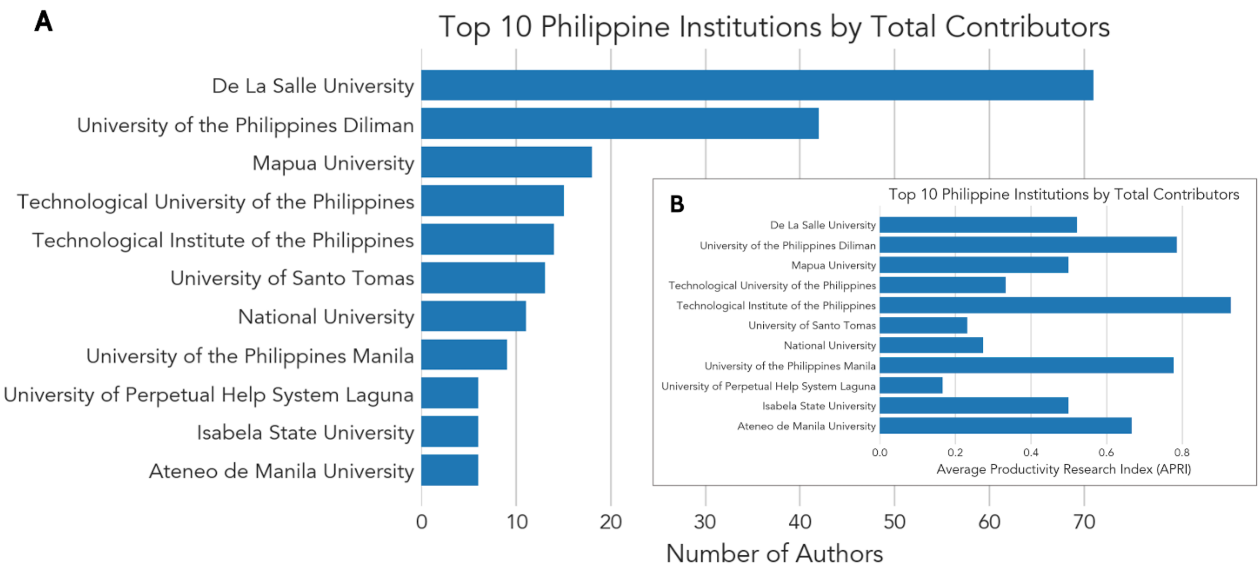

Figure 10: Top 10 Philippine institutions based on the most number of uniquely affiliated authors and the corresponding average productivity research index (APRI).

CONCLUSION

Deep learning is an emerging field in computer science. To ensure that Filipino researchers can actively contribute in this space, we identified the key institutions and researchers working on deep learning in the hope of establishing fundamental collaborations for training, capacity building and institutional strengthening. Analyzing published articles from 2006 to 2020, we found Prospero Naval from UPD as the most productive Filipino researcher working on DL. This is despite having fewer people to collaborate with compared to the more connected/collaborative researchers from DLSU -- Elmer Dadios and Melvin Cabatuan. In terms of collaboration, researchers from UPD, DLSU and UPM were observed to be part of many large research clusters in the DL network, but researchers from DLSU lead the way in working with more institutions. Notably, we have observed that researchers from the two largest research institutions, DLSU and UPD, do not have any history of collaboration on deep learning research.

Disadvantaged by the lack of connectivity, this can pose a challenge for national capacity building and strengthening of DL research. Besides this, the local academic enterprise still relies heavily on undergraduate students to drive research projects, as shown by the many low degree nodes in the network. This practice has been hard to sustain especially for long-term projects. Thus, institutions should continue growing and improving their graduate programs in order to have more mature and research-oriented people who can work on more fundamental problems and make deeper investigations for deep learning.

Among the top 5 most productive institutions, only TIP and UPD have average productivity research index (APRI) greater than 0.6 but less than 1.0. This indicates that the publication output is not well proportioned to the number of Filipino DL researchers in each institution. Although the number of publications does not have one to one correspondence to the number of Filipino researchers in each institution, overall, the Filipino DL research community is doing good work as shown by the increasing trend of works focused on neural and convolutional neural networks since 2017. However, almost all of them are still research that applies deep learning, and not so much on fundamental and or theoretical work that the global research community can directly or indirectly benefit from.

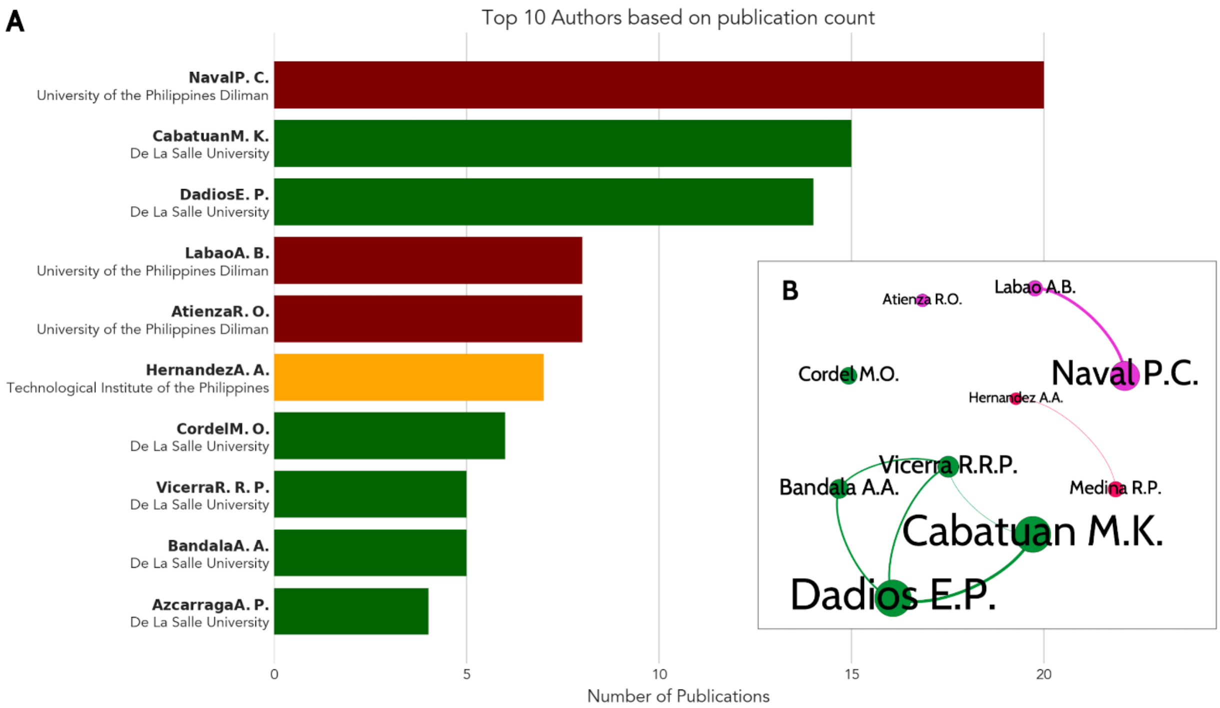

Figure 11: Top 10 authors based on their publication count.

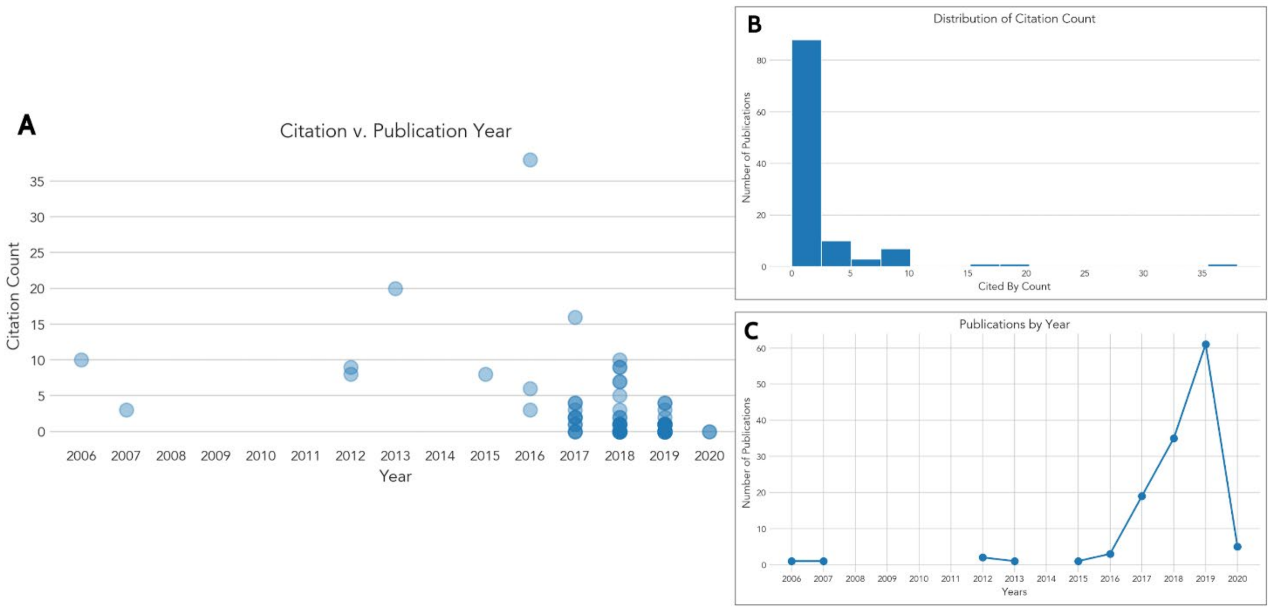

Figure 12: Citation counts of the DL publications per year (until January 2020) and its distribution.

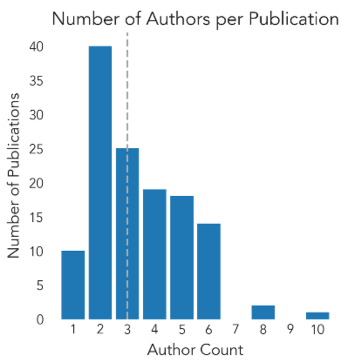

Figure 13: This simply shows how many people work on a single paper. The dotted line represents the median value - 3 authors per paper.

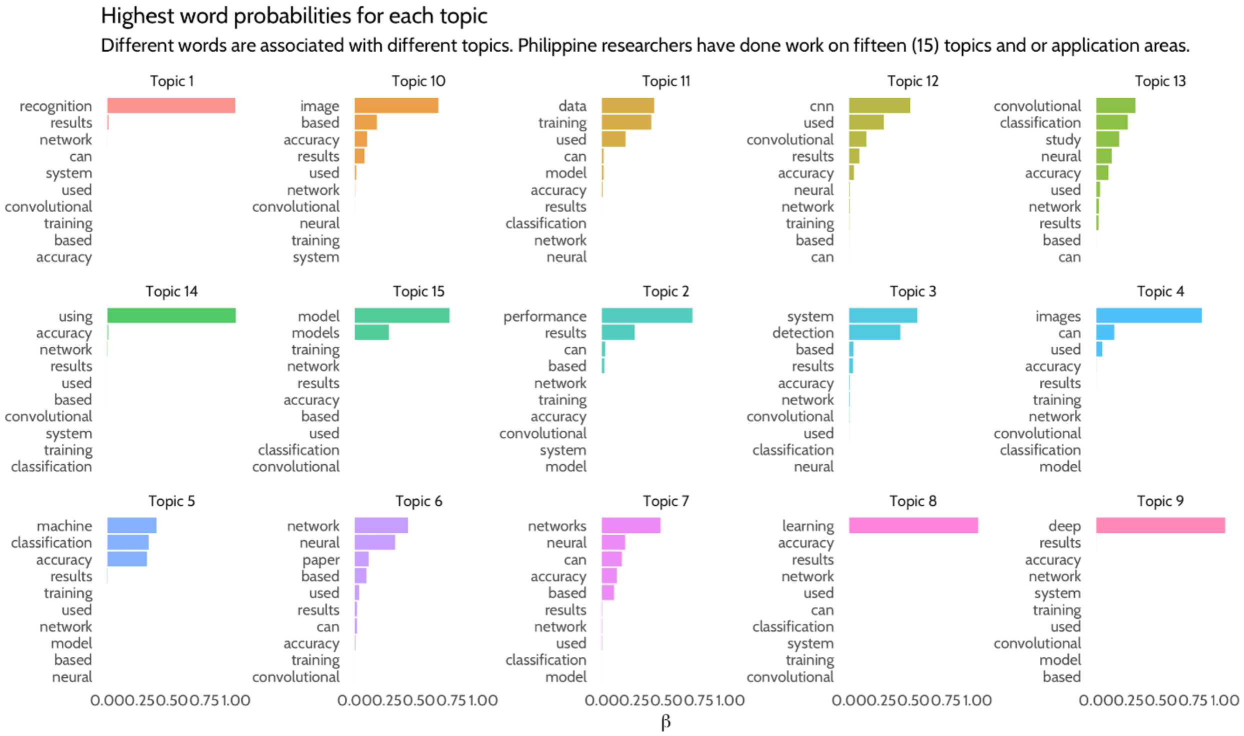

Figure 14: Fifteen identified topics from the Filipino DL publications.

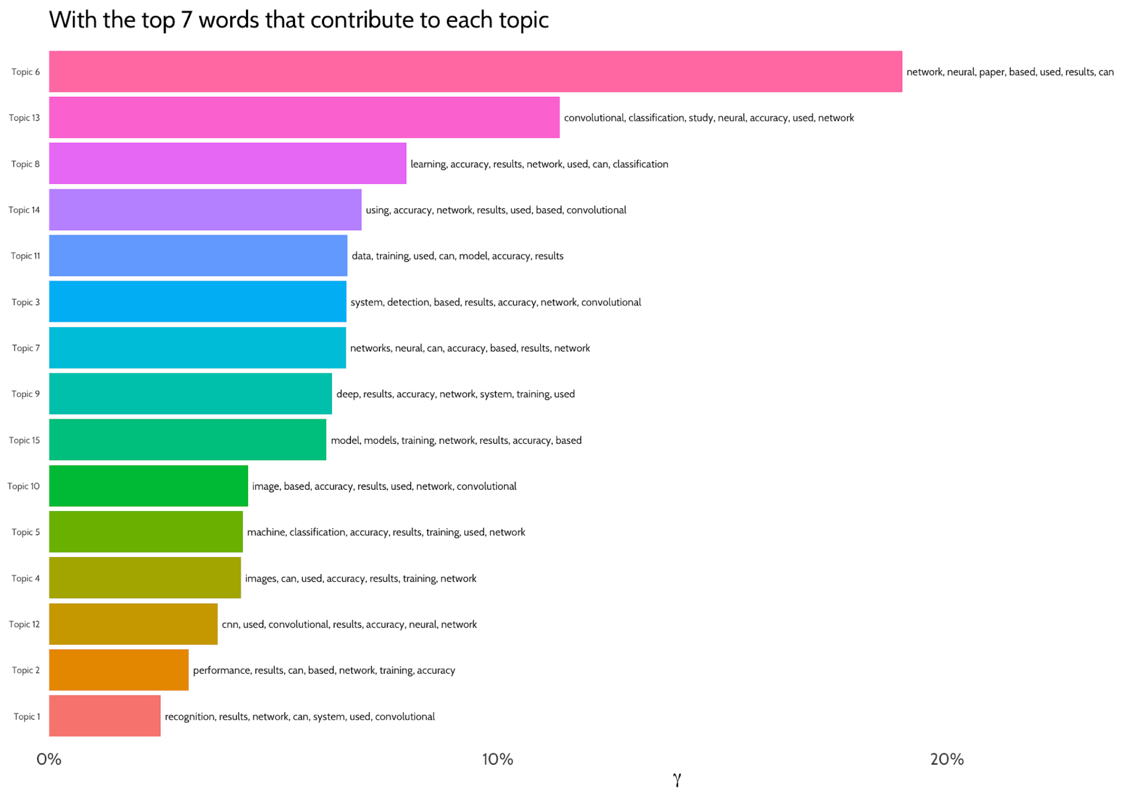

Figure 15: Topics prevalence across DL publications including the top 7 keywords per topic.

The deep learning research community heavily favors conferences over journals in disseminating state-of-the-art works, as is the case in the field of computing and information sciences. While Filipino researchers have long been participating in conferences (around 60% of publications are conference proceedings), most of them are local (Philippine Computing Science Congress) and lower tiered venues. As of writing, only one researcher from DLSU has published a full paper in a top tier conference, which was a collaboration with international counterparts. Thus, each institution has to consider how to support and fund the attendance of researchers in these conferences in order to sustain or increase research outputs in DL. Attending high impact conferences also allow the researchers to network and benchmark their works internationally. But all of these entail a greater push for high quality work that will have generalizable and long-standing contributions to the global community -- something that will be harder to achieve if Filipino DL researchers will continue to focus on producing quick, sometimes niche, applied research works.

To understand the collaboration dynamics or patterns of Filipino researchers, to monitor their current development, and to identify our nation's future research directions, co-authorship network analysis including different fields of scientific research in the country can be performed. This will be beneficial to both Filipino researchers and decision makers who are seeking to improve research productivity of our nation.

ACKNOWLEDGMENTS

AL held a research fellowship from De La Salle University. We would like to thank De La Salle University for the research fellowship funding and Eduardo Mendoza for verifying our methodology and results.

CONFLICTS OF INTEREST

There is no conflict of interest among the authors.

CONTRIBUTIONS OF INDIVIDUAL AUTHORS

AL, UC and BPS collected the data and designed the methodology. UC and BPS performed the analysis and contributed to the writing of the manuscript. AL supervised the work and contributed to the writing and editing of the manuscript.

REFERENCES

Abramo G, D'Angelo CA. How do you define and measure research productivity?. *Scientometrics* 2014; 101(2):1129-44.

Adams J, Loach T. Comment: A well-connected world. *Nature* 2015; 527(7577):S58-9.

Al-Ayyoub M, Arkok B, Jararweh Y. An extended study of collaboration networks of levantine biomedical researchers. *J Ambient Intell Humaniz Comput* 2018; 9(1):3-18.

Brownlee J. Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions. Machine Learning Mastery 2018.

Katz JS, Martin BR. What is research collaboration?. *Res. Policy* 199; 26(1):1-8.

Moolayil JJ. A Layman's Guide to Deep Neural Networks. Medium, Towards Data Science, 25 July 2019, https://towardsdatascience.com/a-laymans-guide-to-deep-neural-networks-ddcea24847fb

Morel CM, Serruya SJ, Penna GO, Guimarães R. Co-authorship network analysis: a powerful tool for strategic planning of research, development and capacity building programs on neglected diseases. *PLoS Neglect Trop D* 2009; 3(8).

Roberts ME, Stewart BM, Tingley D. stm: R package for structural topic models. *J. Stat. Softw.* 2014; 10(2):1-40.

Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* 2019; 9(1):1-2.