Attention: to Better Stand on the Shoulders of Giants

Sha Yuan[1,3†*], Zhou Shao[2,3†], Yu Zhang[4], Tong Xiao[3], and Yifan Wang[3]

† Sha Yuan and Zhou Shao are co-first authors of the article. They contribute equally to this article.

* Sha Yuan is the corresponding author. ✉yuansha@baai.ac.cn

[1] Beijing Academy of Artificial Intelligence

[2] Nanjing University of Science and Technology

[3] Department of Computer Science and Technology, Tsinghua University

[4] Institute of Medical Information, Peking Union Medical College, Chinese Academy of Medical Sciences

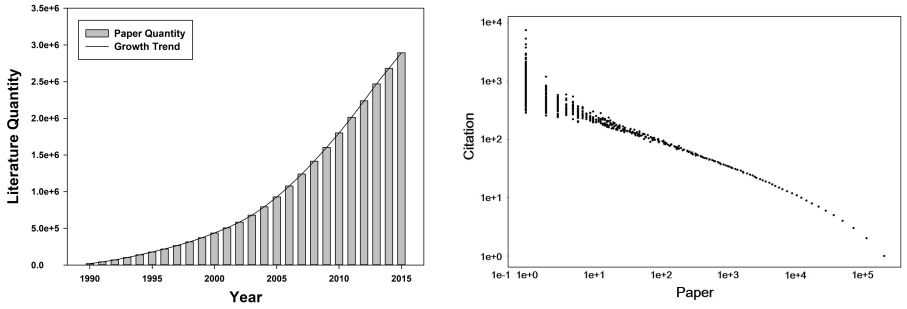
Abstract. Science of Science (SciSci) is an emerging discipline wherein science is used to study the structure and evolution of science itself using large data sets. The increasing availability of digital data on scholarly outcomes offers unprecedented opportunities to explore SciSci. In the progress of science, the previously discovered knowledge principally inspires new scientific ideas, and citation is a reasonably good reflection of this cumulative nature of scientific research. The researchers that choose potentially influential references will have a lead over the emerging publications. Although the peer-review process is the mainly reliable way of predicting a paper's future impact, the ability to foresee the lasting impact based on citation records is increasingly essential in the scientific impact analysis in the era of big data. This paper develops an attention mechanism for the long-term scientific impact prediction and validates the method based on a real large-scale citation data set. The results break conventional thinking. Instead of accurately simulating the original power-law distribution, emphasizing the limited attention can better stand on the shoulders of giants.

Keywords: Science of Science · Scientific Impact · Attention

1 Introduction

With the advent of the era of big data, people pay more and more attention to the value of data. The massive volume of publications created every year has grown into big data that we can't ignore. With the development of SciSci, it provides a quantitative understanding of scientific discovery, creativity, and practice [12, 11, 23, 34, 5, 29]. From the perspective of SciSci, identifying fundamental drivers of science and developing predictive models to capture its evolution are instrumental for successful science. SciSci reveals that the previously discovered knowledge mainly inspires new scientific ideas [47], and citation [24] is a relatively good reflection of this cumulative nature of scientific research. Citation

count, which has been used to evaluate the quality and influence of scientific work for a long time, stands out from many quantification measure metrics of scientific impact [7]. With the rapid evolution of scientific research, there is a massive volume of literature published every year, and this situation is expected to remain within the foreseeable future. Fig. 1 shows the statistics on the citation data set used in this paper. The data set is extracted from AMiner [38], which is a billion-scale academic search and mining system. Fig. 1(a) visualizes the explosive increase in the volume of publications in the past years from 1990 to 2015. It shows that the literature quantity assumes the exponential order to grow.



(a) The volume of literatures. (b) The citation Distribution.

Fig. 1. Statistics of literature data from AMiner.

Scientific work is founded on prior research. It is not wise nor possible for researchers to track all existing related work due to the enormous volume of the existing publications. In general, researchers follow or cite merely a small proportion of high-quality publications. SciSci provides several quantification methods for scientific impact measurement in article-level, author-level, and journal-level. Much SciSci work has been done on the evaluation metrics for the quality and influence of scientific work [31, 30], including citation count [33], h-index [18], and impact factor [16]. One of the most basic quantification measure metrics of scientific impact is citation count. It measures the number of received citations for an article. Many other essential evaluation criteria of authors (e.g., h-index) and journals [15, 13] (e.g., Impact Factor) are calculated based on citation count.

A lot of SciSci researchers have focused on the characterization of scientific impact [35], such as the universal citation distributions [28], the characteristics of citation networks [19, 26, 20], and the growth pattern of scientific impact [10, 17]. The results reveal the regularity of scientific progress that a few research papers attract the vast majority of citations [4], long-distance interdisciplinarity leads to higher scientific impact [21, 46]. Fig. 1(b) illustrates the citation distribution (the number of papers vs. citation counts) of about two million papers. The citation

distribution follows the power-law distribution [14]. It is natural to find that not all publications attract equal attention in academia. A few research papers accumulate the vast majority of citations, and most of the other papers attract only a few citations [4]. A small number of scholarly outcomes are more likely to attract scientists' attention than others accounting for a vast majority. For the ever-growing literature quantity, it is significant to forecast which paper is more likely to attract more attention in the scientific community. Zhu et al. [48] present several machine learning methods and one multiple linear regression strategy to predict a paper's future citation. Ali et al. [1] propose a novel method for predicting long-term citations of a paper based on the number of its citations in the first few years after publication. Daniel et al. [9] present GNN-based architecture that predicts the top set of papers at the time of publication.

The fact is that the current citation count and the derived metrics can only capture the past accomplishment. They lack the predictive power to quantify the future impact [2]. Predicting an individual paper's citation count over time is significant, but (arguably) very difficult. To predict the citation count of individual items within a complex evolving system, current models are falling into two main paradigms. One formulates the citation count over time as time series and then makes predictions by either exploiting temporal correlations [37] or fitting these time series with certain classes of designed functions [25, 3], including the regression models [45], the counting process [39], the point process, the Poisson process [40, 41], Reinforced Poisson Process (RPP) [32], self-excited Hawkes Process [27], RPP with self-excited Hawkes Process [42]. The designed functions consider various factors.

The other prevalent line utilizes Deep Neural Network (DNN) based models to solve the scientific impact prediction problem. Recently, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have received considerable attention from academia and the industry. RNN has been proven to perform particularly well on temporal data series [36]. Due to the vanishing gradient problem, RNN always fails to handle the temporal contingencies present in the input/output sequences spanning long intervals [6]. The networks with loops in RNN allow information to persist for a long time. Long short-term memory (LSTM) is proven to be capable of learning long-term dependencies. RNN with LSTM units performs rather well in handling long-term temporal data series [43].

All the existing methods try to tune the citation distribution precisely as the original power-law distribution. However, this paper argues that the effectiveness of quantifying long-term scientific impact is fundamentally limited in this routine thinking. This paper proposes to put more attention on some specific items, such as highly cited papers. The authors validate the proposed method on a real large-scale citation data set. Extensive experiment results demonstrate that the proposed method possesses remarkable power at predicting long-term scientific citation. The most important contribution is that this paper changes the line of thinking in quantifying the long-term scientific impact. Instead of simulating

the original power-law distribution, researchers need to emphasize the limited attention to better stand on giants' shoulders.

2 Problem Formulation

The primary evaluation metric for scientific impact is citation count. The received citation count of an individual paper d during time period $[0, T]$ is characterized by a time-stamped sequence $\{n_d^t\}_{t=0}^T$, where n_d^t represents the number of citation counts received by paper d at time t, n_d^t is an integer greater than or equal to zero. In giving the historical citation records, the goal is to model the future citation count and predict it over an arbitrary time.

Given the literature corpus D, $\text{card}(D) = M$ means find M papers from D. In this paper, we believe that the scientific impact of a literature article is equal to the number of papers which cite it. The scientific impact of a literature article $d \in D$ at time t is defined as its citation counts n_d^t:

$$citing_d^t = \{\tilde{d} \in D, \tilde{d} \neq d : \tilde{d}^t \ cites \ d\}, n_d^t \quad = \text{card}(citing_d^t). \tag{1}$$

The underlying assumption of the citation count here is the accumulated citations, which make it possible to quantify citations for different items at different times. The long-term scientific impact of the individual item d can be formalized as the following time series $\{n_d^0, \cdots, n_d^t, \cdots, n_d^T\}$. Without loss of generality, the number of accumulated citation counts increases over time. And then, we have $0 = n_d^0 \leq \cdots \leq n_d^t \leq \cdots \leq n_d^T = N_d$.

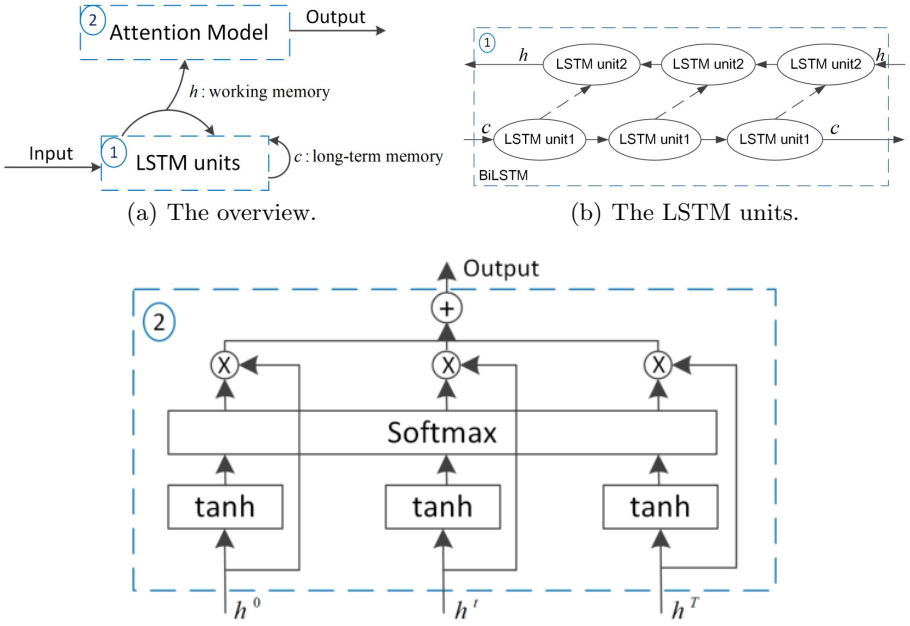
In the scientific impact prediction problem, the input $\vec{X}$ is $\{n_d^0, \cdots, n_d^t, \cdots\}$, for every paper $d \in D$, where n_d^t is the citation counts of paper d at time t. The goal of the scientific impact prediction problem is to learn a predictive function f to predict the citation counts of an article d after a given period time t. Formally, we have:

$$f(d|\vec{X}, t) \rightarrow \hat{n}_d^t, \tag{2}$$

where $\hat{n}_d^t$ is the predicted citation count and n_d^t is the actual one. Based on the learned prediction function, we can predict the citation count of a paper for the next years. For example, the citation count of paper d at time t is given by $f(d|\vec{X}, t)$.

3 Scientific Impact Prediction

As the most efficient scientific impact prediction method found so far, RNN has already achieved compelling performance in predicting the scientific impact. This paper embeds the RNN with LSTM units as a baseline and then emphasizes highly cited articles in the proposed attention mechanism. Although many other fields have used the attention mechanism, the proposed method gives new insight into long-term quantifying scientific impact. Instead of adapting citation distribution to a power-law distribution, this paper's findings provide a new line of thinking for the SciSci research.



(a) The overview. (b) The LSTM units.

(c) The attention model.

Fig. 2. The deep learning attention model.

3.1 Deep Learning Attention Mechanism

Given a time-stamped sequence $\{n_d^t\}_{t=0}^T$, a K-dimensional feature vector $\vec{X} = \{x_d^0, \cdots, x_d^t, \cdots, x_d^T\}$ needs to be designed as input. The input space of every item with popularity records $\{(x^0, n^0), \cdots, (x^t, n^t), \cdots, (x^T, n^T)\}$ reflects the intrinsic quality of the item. Fig. 2(a) gives an overview of the model architecture. There are two critical components in the architecture: the RNN with LSTM units and the attention model. As illustrated in Fig. 2(b), it arranges the LSTM units in the form of RNN with L layers. In the deep neural network, the parameter L depends on the input scale. RNN is famous for its popularity and well-known capability for efficient time series learning [43]. The LSTM units capture the long-range dependency in long-term scientific impact quantification.

The RNN with LSTM Units. The LSTM units are arranged in the form of RNN, as illustrated in Fig. 2(b). There are four major components in a standard LSTM unit, including a memory cell, a forget gate Γ_f, an input gate Γ_i, and an output gate Γ_o. The gates are responsible for information processing and storage over arbitrary time intervals. Usually, the outputs of these gates are between 0 and 1. A new study gives suggestions to push the output values of the gates towards 0 or 1. By doing so, the gates are mostly open or closed instead of in a middle state [22]. This paper arranges the LSTM units in the form of RNN. In

this way, introducing the memory cell will solve the vanishing gradient problem. Thus, it can store information for either short or long periods in the LSTM unit.

Intuitively, the input gate controls the extent to which a new value flows into the memory cell. The input gate's function passes through the input gate and is added to the cell state to update it. The following formula for the input gate is used:

$$\Gamma_i^t = \sigma\left(W_i\left[h^{t-1}, x^t\right] + b_i\right), \tag{3}$$

where matrix W_i collects the weights of the input and recurrent connections. The symbol σ represents the Sigmoid function. The values of the vector Γ_i^t are between 0 and 1. If one of the values of Γ_i^t is 0 (or close to 0), it means that this input gate is closed, and no new information is allowed into the memory cell at time t. If one of the values is 1, the input gate is open for a new coming value at time t. Otherwise, the gate is in the state of half-open half-clearance.

The forget gate controls the extent to which a value remains in the memory cell. It provides a way to get rid of the previously stored memory value. Here is the formulation of the forget gate:

$$\Gamma_f^t = \sigma\left(W_f\left[h^{t-1}, x^t\right] + b_f\right), \tag{4}$$

where W_f is the weight matrix that governs the behavior of the forget gate. Similar to Γ_i^t, Γ_f^t is also a vector of values between 0 and 1. If one of the values of Γ_f^t is 0 (or close to 0), it means that the memory cell should remove that piece of information in the corresponding component in the cell. If one of the values is 1, the corresponding information will be kept.

Remembering information for long periods is practically the default behavior of LSTM. The long-term accumulative influence is formulated as follows:

$$c^t = \Gamma_f^t * c^{t-1} + \Gamma_i^t * \tilde{c}^t, \tag{5}$$

where $*$ denotes the Hadamard product (the element-wise multiplication of matrices), $\tilde{c}^t$ is calculated as follows:

$$\tilde{c}^t = \tanh\left(W_c\left[h^{t-1}, x^t\right] + b_c\right). \tag{6}$$

That is, the information in the memory cell consists of two parts: the retained old information $\Gamma_f^t * c^{t-1}$ (controlled by the forget gate) and the new coming information $\Gamma_i^t * \tilde{c}^t$ (controlled by the input gate).

The output gate controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit. The following output function is used:

$$\Gamma_o^t = \sigma\left(W_o\left[h^{t-1}, x^t\right] + b_o\right). \tag{7}$$

The weight matrices and bias vector parameters are needed to be learned during training. This paper updates the current working state as the following formula:

$$h^t = \Gamma_o^t * \tanh\left(c^t\right). \tag{8}$$

The items stored in the current working state have an advantage in reading over those stored in long-term memory. In the time series modeling of scientific impact, the recent items stored in the short-term working state have an advantage over those stored in the long-term memory. The next step introduces the attention mechanism based on h^t.

The Attention Model. The artificial attention mechanism, inspired by the attention behavior in neuroscience, has been applied in deep learning for speech recognition, translation, and visual identification of objects.Broadly, attention mechanisms are components of prediction systems that allow the system to focus on different subsets of the input sequentially. It aims to capture the critical points and focuses on the relevant parts more than the remote parts as a human does. More specifically, content-based attention generates attention distribution. Only part of a subset of the input information is focused. The attention function needs to be differentiable, so that everywhere of the input is focused, just to different extents.

The deep learning attention mechanism used in this paper works as follows: given an input $\vec{X} = \{x_d^0, \cdots, x_d^t, \cdots, x_d^T\}$, the aforementioned LSTM units generate $\vec{h} = \{h_1, \cdots, h_t, \cdots, h_T\}$ to represent the hidden patterns of the input. The output is the summary of the h_t focusing on information linked to the input. In this formulation, attention produces a fixed-length embedding of the input sequence by computing an adaptive weighted average of the state sequence $\vec{h}$.

The graphical representation of the attention model is shown in Fig. 2(c). The input $\vec{X}$ and the hidden layer $\vec{h}$ of the LSTM network (an RNN composed of LSTM units) are the input of the attention model. Then, it computes the following formula:

$$a^t = \tanh\left(W_a \left[x^t, h^t\right]\right), \tag{9}$$

where W_a is the weight matrix. An important remark here is that each a^t is computed independently without looking at the other $x^{t'}$ for $t' \neq t$. Then, each a^t is linked to a Softmax layer, which function is given by:

$$\alpha^t = \frac{e^{a^t}}{\sum_t e^{a^t}}, \text{for } t = 1, \cdots, T \tag{10}$$

where $\sum_t \alpha^t = 1$, the α^t is the softmax of the a^t projected on a learned direction. The output is a weighted arithmetic mean of the input, and the weights reflect the relevance of $\vec{h}$ and the input. It is calculated as the following formula:

$$O = \sum_t \alpha^t x_t. \tag{11}$$

Finally, the popularity of item d at time t is given by the prediction $f(d|\vec{X}, t) = O$.

3.2 Key Factor in Quantifying Long-term Impact

As widely acknowledged, the citation distribution follows the power-law distribution. This finding leads the way for research in this domain. Researchers try

to simulate the citation distribution as the power-law distribution. This paper changes the line of thinking. Although the number of research papers has exploded, the reading time of scientists has not. At the same time, the attention shifts toward the top 1% over time [4]. Even though the citation distribution follows the power-law distribution, attention is vital in quantifying the long-term scientific impact.

In the fact of limited attention, the Matthew effect dominates in quantifying the long-term scientific impact. The experiments will confirm it. The citation count captures the inherent differences between papers, accounting for the perceived novelty and the importance of a paper. The "rich-get-richer" phenomenon summarizes the Matthew effect of accumulated advantage, i.e., previously accumulated attention triggers more subsequent attention [8] than others. In fact, the highly popular items are more visible and more likely to be viewed than others. The proposed model emphasizes highly cited papers under limited attention. The memory cell in the LSTM unit considers the long-term dependencies. As shown in Eq. (5), previously accumulated attention stored in the long-term memory triggers more subsequent attention. What is more, the attention model, which focuses on the most popular part of the time series as Eq. (11) does, also emphasizes the Matthew effect.

4 Experiments

This section demonstrates the effectiveness of putting particular emphasis on the vital factor in quantifying the long-term scientific impact.

4.1 Dataset

The authors extract the data from an academic search and mining platform called AMiner and construct a real large-scale scholarly dataset–Academic Social Network[1]. The citation network's full graph in this dataset has about 2 million vertices (papers) and 8 million edges (citations). In detail, the dataset is composed of $2,092,356$ digitalized papers spanning from 1936 to 2016 (for more than 80 years) and $8,024,869$ citations between them. By convention, the authors eliminate those papers with less than 5 citations during the first 5 years after publication and only retain the remaining papers as the training data. As a result, $143,902$ papers published from 1956 to 2015 are retained.

4.2 Baseline Models and Evaluation Metrics

To compare the predictive performance of the proposed attention model against other models, we introduce several published models that have been used to predict scientific impact. Specifically, the experiments' comparison methods are LR, CART, SVR (the three basic machine learning methods used in [45]), RPP [40,

[1] https://www.aminer.cn/data

32], and RNN [43]. The advantage of deep learning is the utilization of various features. For fairness, the authors only use the citation count records and the same feature used in [40, 32]. For the fair comparison among different kinds of models, all models use the same vector features, which are the citation records of the first five years after publication. Statistics show that for most papers, the first five years after publication can well reflect their influence on the current research stage.

This paper uses two basic scientific impact evaluation metrics: Mean Absolute Percentage Error (MAPE) and Accuracy (ACC). Let n_d^t be the observed citations of paper d up to time t, and $\hat{n}_d^t$ be the predicted one. The MAPE measures the average deviation between the predicted and observed citations over all the papers. For a dataset of M papers, the MAPE is given by:

$$\text{MAPE} = \frac{1}{M}\sum_{d=1}^{M}\left|\frac{\hat{n}_d^t - n_d^t}{n_d^t}\right|. \tag{12}$$

ACC measures the fraction of papers correctly predicted under a given error tolerance ϵ. Specifically, the accuracy of citation prediction over M papers is defined as:

$$\text{ACC} = \frac{1}{M}\sum_{d=1}^{M}\text{I}\left[\left|\frac{\hat{n}_d^t - n_d^t}{n_d^t}\right| \leq \epsilon\right], \tag{13}$$

where $\text{I}[\theta]$ is an indicator function which returns 1 if the statement θ is true, otherwise returns 0. We find that our method always outperforms regardless of ϵ's value. In this paper, we set $\epsilon = 0.3$.

4.3 Model Setting

The experiment results show that the longer the duration of the training set, the better the long-term prediction performance. According to our experiment, this paper sets the training period as 5 years and then predicts the citation counts for each paper from 1^{st} to 5^{th} after the training period. For example, $t = 1$ means that the first observation year after the training period. In the experiments, the features with positive contributions are the citation history, the current h-index of the paper author, and the publication journal level [44]. For the convenience of performance comparison, the input feature used here is the citation history for every sub-window of length 10 years. The value of the parameter L is 2. The loss function used here is MAPE. Adadelta is the gradient descent optimization algorithm. The attention layer is fully connected and uses tanh activation. And the code is available on github [2].

4.4 Results

As shown in Table. 1, the proposed model exhibits the best performance in terms of ACC in all the situations of t = 1, 2, 3, 4, and 5. It means that the

[2] https://github.com/AIOpenData/attention

DLAM consistently achieves higher accuracy than other models across different observation times. What is more, the proposed model also exhibits the best performance in terms of MAPE in all the situations mentioned above. That is, the proposed model achieves higher accuracy and lowers error rates simultaneously. In the experiments, all the models used for comparison achieve acceptable low error rates, except RPP. RPP can avoid this problem with prior [32], which incorporates conjugate prior for the fitness parameter. However, the RPP with prior does not improve the ACC performance. Overall, the proposed model also outperforms RPP with prior.

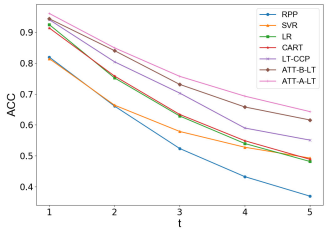
Table 1. The performance of various models on the data set.

	$t=1$		$t=2$		$t=3$		$t=4$		$t=5$	
Models	MAPE	ACC	MAPE	ACC	MAPE	ACC	MAPE	ACC	MAPE	ACC
RPP	0.219	0.819	0.381	0.661	0.686	0.524	0.904	0.433	1.376	0.370
SVR	0.195	0.814	0.252	0.664	0.296	0.579	0.331	0.528	0.362	0.493
LR	0.136	0.924	0.207	0.752	0.269	0.629	0.330	0.540	0.386	0.482
CART	0.131	0.913	0.202	0.758	0.256	0.634	0.297	0.549	0.328	0.489
RNN	0.123	0.940	0.185	0.804	0.234	0.703	0.298	0.590	0.317	0.551
DLAM	**0.121**	**0.960**	**0.168**	**0.849**	**0.203**	**0.757**	**0.231**	**0.693**	**0.255**	**0.643**

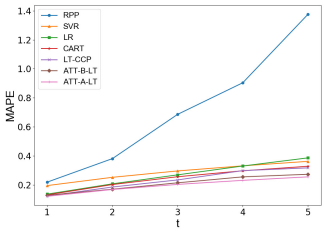
Compared to the other methods in terms of ACC and MAPE, the proposed model increases with the number of years after the training period. Compare to RNN (the most efficient method certified in recent works), the proposed model achieves a few performance improvements, about 1.65% in terms of MAPE and about 2.13% in terms of ACC when $t = 1$. However, in the situation of $t = 5$, the proposed model achieves significant performance improvement of about 24.31% in terms of MAPE and about 16.7% in terms of ACC. In other words, the proposed model shows much superiority over other models in scientific impact prediction, especially in the **long-term** situation.

5 Further Exploration

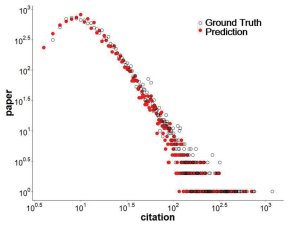
Effectiveness of the attention mechanism. The authors remove the attention module of the proposed model to verify the effectiveness of the attention mechanism. The remainder is RNN with LSTM units (labeled as LT-CCP), proven to be useful in long-term citation count prediction. In the next step, we add the attention mechanism in two different ways. Firstly, we add the attention module before the RNN module, labeled as ATT-B-LT (attention before LT-CCP). In a second way, we add the attention module after the RNN module, labeled as ATT-A-LT (attention after LT-CCP). As shown in Fig. 3(b) and Fig. 3(a), the ACC is increased, and the corresponding MAPE is decreased. Both ATT-B-LT and ATT-A-LT perform better than LT-CCP in terms of MAPE and ACC. Introducing the attention module improves the ability of scientific impact prediction. The effectiveness of the attention mechanism is verified.



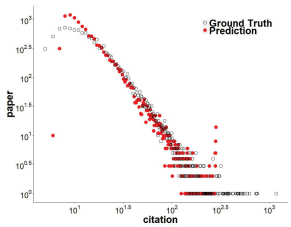
(a) ACC comparision.



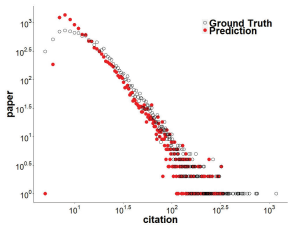
(b) MAPE comparision.



(c) LT-CCP (RNN with LSTM).



(d) ATT-B-LT.



(e) ATT-A-LT (DLAM).

Fig. 3. The performance comparison in citation count prediction.

In addition, we can see that the ATT-A-LT performs better than ATT-B-LT. When the attention model is applied after a deep learning model, it is more effective than the reverse combination. It indicates that the deep learning model can learn the implicit features underlying the citation records, which boots the performance.

Analysis of the citation distribution. We illustrate the actual and the predicted citation distribution of LT-CCP (RNN with LSTM), ATT-B-LT, and ATT-A-LT (DLAM) when $t = 5$ in Fig. 3(c), Fig. 3(d), and Fig. 3(e), respectively. The LT-CCP (RNN with LSTM) illustrated in Fig. 3(c) shows the best simulation of the power-law distribution. But the ATT-B-LT shown in Fig. 3(d) and the ATT-A-LT (DLAM) shown in Fig. 3(e) present bad simulation of the power-law distribution. The results show that LT-CCP (RNN with LSTM) matches very well with that of real citations, but the ATT-B-LT and the ATT-A-LT (DLAM) don't. Usually, it is believed that the more similar the power-law distribution, the whole result is better. At first glance, it seems that LT-CCP (RNN with LSTM) performs the best.

However, the first thought is wrong. As verified in Fig. 3(a) and Fig. 3(b), the LT-CCP (RNN with LSTM) performs the worst. In fact, the LT-CCP only has a better fitting effect on the papers with little citation counts. On the contrary, the ATT-B-LT and ATT-A-LT (DLAM) have a better fitting effect on the highly cited papers. The methods with attention mechanisms achieve better overall performance than others. It is more accordant with practical prediction requirements that a few papers occupy a vast number of citations. It further proves the

effectiveness of the attention model. The experimental results indicate that we need to change the fixed pattern of thinking in quantifying long-term scientific impact.

6 Conclusion

Scientific impact evaluation is always a critical point in decision-making concerning recruitment and funding in the scientific community. The rapid evolution of scientific research has been creating a massive volume of publications every year. Among the many quantification measures of scientific impact, citation count stands out for its frequent use in the research community. Although the peer-review process is the mainly reliable way of predicting a paper's future impact, the ability to foresee the lasting impact based on citation records is increasingly essential in the scientific impact analysis in the era of big data.

SciSci provides a quantitative understanding of the scientific impact based on big data empirical analysis. In this paper, the authors develop an attention mechanism in long-term scientific impact prediction and verify its effectiveness. More importantly, this paper provides us great insights into understanding the critical factor in quantifying the long-term scientific impact. Usually, researchers try to make the predicted citation distribution similar to the original one. However, the experimental results in the paper question this solution. In future research work, we need to change the fixed pattern of thinking in quantifying the long-term scientific impact and emphasize limited attention to better stand on giants' shoulders.

Acknowledgement

The work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61806111 and NSFC for Distinguished Young Scholar under Grant No. 61825602.

References

1. Abrishami, A., Aliakbary, S.: Predicting citation counts based on deep neural network learning techniques. Journal of Informetrics **13**(2), 485–499 (2019)
2. Acuna, D.E., Allesina, S., Kording, K.P.: Future impact: Predicting scientific success. Nature **489**(7415), 201 (2012)
3. Bao, P., Shen, H.W., Huang, J., Cheng, X.Q.: Popularity prediction in microblogging network: a case study on sina weibo. In: International Conference on World Wide Web. pp. 177–178 (2013)
4. Barabási, A.L., Song, C., Wang, D.: Publishing: Handful of papers dominates citation. Nature **491**(7422), 40 (2012)
5. Barbosa, S.D.J., Silveira, M.S., Gasparini, I.: What publications metadata tell us about the evolution of a scientific community: the case of the brazilian human–computer interaction conference series. Scientometrics **110**(1), 275–300 (2017)

6. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks **5**(2), 157–166 (1994)
7. Cao, X., Chen, Y., Liu, K.R.: A data analytic approach to quantifying scientific impact. Journal of Informetrics **10**(2), 471–484 (2016)
8. Crane, R., Sornette, D.: Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences of the United States of America **105**(41), 15649–53 (2008)
9. Cummings, D., Nassar, M.: Structured citation trend prediction using graph neural networks. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3897–3901. IEEE (2020)
10. Dong, Y., Johnson, R.A., Chawla, N.V.: Will this paper increase your h-index?: Scientific impact prediction. In: Proceedings of the eighth ACM international conference on web search and data mining. pp. 149–158 (2015)
11. Dong, Y., Ma, H., Shen, Z., Wang, K.: A century of science: Globalization of scientific collaborations, citations, and innovations. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1437–1446. ACM (2017)
12. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., et al.: Science of science. Science **359**(6379) (2018)
13. Franceschet, M.: The role of conference publications in cs. Communications of the ACM **53**(12), 129–132 (2010)
14. Frank, M.R., Wang, D., Cebrian, M., Rahwan, I.: The evolution of citation graphs in artificial intelligence research. Nature Machine Intelligence **1**(2), 79 (2019)
15. Freyne, J., Coyle, L., Smyth, B., Cunningham, P.: Relative status of journal and conference publications in computer science. Communications of the ACM **53**(11), 124–132 (2010)
16. Garfield, E.: Impact factors, and why they won't go away. Nature **411**(6837), 522 (2001)
17. Greene, M.: The demise of the lone author. Nature **450**(7173), 1165 (2007)
18. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America **102**(46), 16569–16572 (2005)
19. Hunter, D., Smyth, P., Vu, D.Q., Asuncion, A.U.: Dynamic egocentric models for citation networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 857–864 (2011)
20. Kuhn, T., Perc, M., Helbing, D.: Inheritance patterns in citation networks reveal scientific memes. Physical Review X **4**(4), 041036 (2014)
21. Larivière, V., Haustein, S., Börner, K.: Long-distance interdisciplinarity leads to higher scientific impact. Plos one **10**(3) (2015)
22. Li, Z., He, D., Tian, F., Chen, W., Qin, T., Wang, L., Liu, T.Y.: Towards binary-valued gates for robust lstm training. In: International Conference on Machine Learning. pp. 3001–3010 (2018)
23. Ma, Y., Uzzi, B.: Scientific prize network predicts who pushes the boundaries of science. Proceedings of the National Academy of Sciences **115**(50), 12608–12615 (2018)
24. Margolis, J.: Citation indexing and evaluation of scientific papers. Science **155**(3767), 1213–1219 (1967)
25. Matsubara, Y., Sakurai, Y., Prakash, B.A., Li, L., Faloutsos, C.: Rise and fall patterns of information diffusion: model and implications. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 6–14 (2012)

26. Pan, R.K., Kaski, K., Fortunato, S.: World citation and collaboration networks: uncovering the role of geography in science. Scientific reports **2**, 902 (2012)
27. Peng, B.: Modeling and predicting popularity dynamics via an influence-based self-excited hawkes process. In: ACM International on Conference on Information and Knowledge Management. pp. 1897–1900 (2016)
28. Radicchi, F., Fortunato, S., Castellano, C.: Universality of citation distributions: Toward an objective measure of scientific impact. Proceedings of the National Academy of Sciences **105**(45), 17268–17272 (2008)
29. Rzhetsky, A., Foster, J.G., Foster, I.T., Evans, J.A.: Choosing experiments to accelerate collective discovery. Proceedings of the National Academy of Sciences **112**(47), 14569–14574 (2015)
30. Sekercioglu, C.H.: Quantifying coauthor contributions. Science **322**(5900), 371–371 (2008)
31. Shen, H.W., Barabási, A.L.: Collective credit allocation in science. Proceedings of the National Academy of Sciences **111**(34), 12325–12330 (2014)
32. Shen, H., Wang, D., Song, C., Barabási, A.L.: Modeling and predicting popularity dynamics via reinforced poisson processes. In: Twenty-eighth AAAI conference on artificial intelligence (2014)
33. Shen, Z., Yang, L., Wu, J.: Lognormal distribution of citation counts is the reason for the relation between impact factors and citation success index. Journal of Informetrics **12**(1), 153–157 (2018)
34. Sinatra, R., Deville, P., Szell, M., Wang, D., Barabási, A.L.: A century of physics. Nature Physics **11**(10), 791 (2015)
35. Sinatra, R., Wang, D., Deville, P., Song, C., Barabási, A.L.: Quantifying the evolution of individual scientific impact. Science **354**(6312), aaf5239 (2016)
36. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112 (2014)
37. Szabo, G., Huberman, B.A.: Predicting the popularity of online content, vol. 53. Communications of the ACM (2010)
38. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer:extraction and mining of academic social networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 990–998 (2008)
39. Vu, D.Q., Asuncion, A.U., Hunter, D.R., Smyth, P.: Dynamic egocentric models for citation networks. In: International Conference on International Conference on Machine Learning. pp. 857–864 (2011)
40. Wang, D., Song, C., Barabási, A.L.: Quantifying long-term scientific impact. Science **342**(6154), 127–132 (2013)
41. Wang, J., Mei, Y., Hicks, D.: Comment on "quantifying long-term scientific impact". Science **345**(6193), 149–149 (2014)
42. Xiao, S., Yan, J., Li, C., Jin, B.: On modeling and predicting individual paper citation count over time. In: Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16). pp. 2676–2682 (2016)
43. Xiao, S., Yan, J., Yang, X., Zha, H., Chu, S.M.: Modeling the intensity function of point process via recurrent neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence. vol. 17, pp. 1597–1603 (2017)
44. Yan, R., Huang, C., Tang, J., Zhang, Y., Li, X.: To better stand on the shoulder of giants. In: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries. pp. 51–60 (2012)

45. Yan, R., Tang, J., Liu, X., Shan, D., Li, X.: Citation count prediction: learning to estimate future citations for literature. In: ACM International Conference on Information and Knowledge Management. pp. 1247–1252 (2011)
46. Yegros-Yegros, A., Rafols, I., D'Este, P.: Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. PloS one **10**(8) (2015)
47. Yin, Y., Wang, D.: The time dimension of science: Connecting the past to the future. Journal of Informetrics **11**(2), 608–621 (2017)
48. Zhu, X.P., Ban, Z.: Citation count prediction based on academic network features. In: 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA). pp. 534–541. IEEE (2018)