Realistic Citation Count Prediction Task for Newly Published Papers

Jun Hirako **Ryohei Sasano** **Koichi Takeda**
Graduate School of Informatics, Nagoya University
hirako.jun.e5@s.mail.nagoya-u.ac.jp
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

Citation count prediction is the task of predicting the future citation counts of academic papers, which is particularly useful for estimating the future impacts of an ever-growing number of academic papers. Although there have been many studies on citation count prediction, they are not applicable to predicting the citation counts of newly published papers, because they assume the availability of future citation counts for papers that have not had enough time pass since publication. In this paper, we first identify problems in the settings of existing studies and introduce a realistic citation count prediction task that strictly uses information available at the time of a target paper's publication. For realistic citation count prediction, we then propose two methods to leverage the citation counts of papers shortly after publication. Through experiments using papers collected from arXiv and bioRxiv, we demonstrate that our methods considerably improve the performance of citation count prediction for newly published papers in a realistic setting.

1 Introduction

In recent years, the number of academic papers in various fields has increased drastically. Accordingly, the demand for techniques for predicting papers that will become influential in the future is growing to help readers identify those papers and support efficient knowledge acquisition. In this study, we adopt the citation count as a measure of future impact, following several previous studies (e.g., Chubin and Garfield, 1979; Aksnes, 2006), and we address the citation count prediction task, which entails predicting how many times a target paper will be cited in the future.

There have been many studies on citation count prediction (e.g., Fu and Aliferis, 2008; van Dongen et al., 2020). However, none of those settings is strictly applicable to predicting the citation count of newly published papers, because they assume the availability of future citation counts for papers shortly after publication. For example, consider the case of predicting the citation counts one year after publication. For training and testing, the correct citation count of the target paper one year after publication must be known; hence, only papers published more than one year ago are used in the experiments. Consequently, even for papers published less than one year before the target paper, the number of citations one year after publication is available, and these citation counts are commonly used to train the prediction model. The bars (■+□) in Figure 1 represent the citation count information used in such settings. However, in actually predicting the future citation count of newly published papers, the correct citation counts one year after the publication of papers published less than one year ago are not available; what is actually available is the gray part of each bar (■) in the figure.

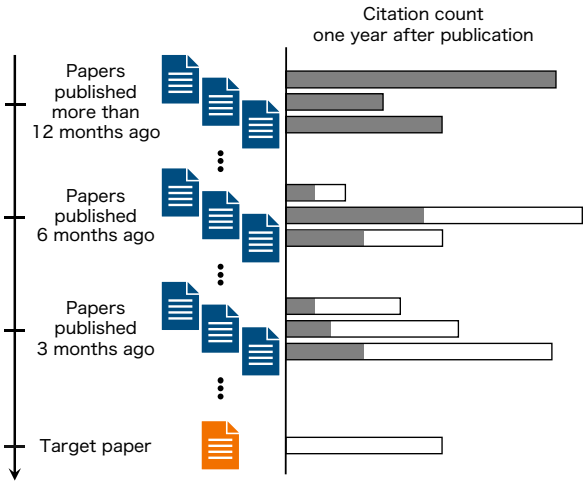

Figure 1: Comparison of a realistic citation count prediction setting with existing research settings. Each bar (■+□) represents the citation count one year after publication, which existing studies assume to be available, while the gray part (■) represents the citation count that is actually available at the time of a target paper's publication.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1131–1141
May 2-6, 2023

The unrealistic assumption in previous studies might appear to have a limited impact on the performance of a prediction model. However, information on the future citation counts of recently published papers could cause leakage of research trends in the near future, which turns out to have a non-negligible impact on performance. Hence, in this study, we first show that the settings of existing studies leak future information that contributes significantly to the prediction performance. We then introduce a realistic citation count prediction task that strictly uses information available at the time of a target paper's publication.

Furthermore, we propose two methods to capture research trends in the near future that are applicable even in our realistic setting. The first method is **citation count complementation**, which uses papers published less than one year ago as training data by estimating the citation count one year after publication from the current citation count. The second method leverages the **degree of early adoption** by using the property that papers that cite more recent papers and papers that cite more frequently cited papers tend to receive more attention in the future.

2 Datasets

For the experiments here, we used two datasets: a CL dataset, consisting of papers in the field of computational linguistics, and a Bio dataset, consisting of papers in the field of biology.

To construct the **CL dataset**, we collected 16,940 papers submitted to arXiv in the Computation and Language (cs.CL) category[1] from June 2014 to June 2020. We considered preprints suitable for this study because they include papers that have not been peer-reviewed and are expected to have a large variance in their future impact. We then obtained the publication dates of papers that cited the collected papers from Semantic Scholar[2] to calculate the citation count for each elapsed month after the publication of each paper in the dataset.

We created 13 subsets, each of which consists of papers published in one of the months from June 2019 to June 2020 and papers published in the five years prior to that month. Within each subset, the papers published in the latest month were used for evaluation, and the remainder was used for training. For example, one subset consists of papers published from May 2015 to May 2020, of which papers published in May 2020 were used for evaluation and the remainder for training. The subsets created in this way have the same properties as cross-validation, where there is overlap in the papers for training, but the papers for evaluation are completely different. The average numbers of papers per subset for training and evaluation are 13,227 and 500.2, respectively. In the experiments, we used the subset that used papers published in June 2019 for evaluation as the development set and the remaining 12 subsets to train and evaluate the model.

To construct the **Bio dataset**, we collected 7,535 papers submitted to the Biochemistry and Plant Biology, Pharmacology and Toxicology areas of bioRxiv[3] from May 2015 to April 2021. As with the CL dataset, we created 12 subsets with papers published in each month from May 2020 to April 2021 as the papers for evaluation. The average numbers of papers per subset for training and evaluation were 5,913 and 257, respectively.[4]

3 Task Formulation

3.1 Leakage in Existing Settings

Most previous studies on citation count prediction adopted the citation count n years after publication as the target citation count for prediction (e.g., Fu and Aliferis, 2008; van Dongen et al., 2020). Those studies used datasets consisting of papers published in a specific time period. Specifically, they used a set of newly published papers by year or a set of randomly selected papers as the evaluation set, and the rest as the training set. The citation count prediction model was then trained using the citation counts n years after the publication of each paper in the training set, and the prediction performance was evaluated by predicting the citation counts of the papers in the evaluation set.

In reality, the citation counts n years after publication are available only for papers published more than n years after publication, but existing settings use those citation counts even for papers published less than n years after publication (Fu and Aliferis, 2008; Davletov et al., 2014; Singh et al., 2015; Abrishami and Aliakbary, 2019; van Dongen et al., 2020). The use of future citation counts that are not actually available in the existing settings may lead

[1] https://arxiv.org/list/cs.CL/recent
[2] https://www.semanticscholar.org/
[3] https://www.biorxiv.org/
[4] Statistics for each subset of the two datasets are provided in Appendix A.

to leakage of future research trends. Accordingly, we conducted a preliminary experiment to examine the effect of this leakage. We found that, with the same number of papers used for training, the use of future citation counts of newly published papers, which are not actually available, achieves higher performance than the use of papers published more than n years ago.[5] Hence, we introduce a realistic citation count prediction task that prevents such leakage and is applicable to the prediction of citation counts for newly published papers.

3.2 Realistic Citation Count Prediction

Our realistic citation count prediction task restricts the citation count information used for training to information that is strictly available as of the publication of the target papers for evaluation. Specifically, in the case of predicting the citation count n years after publication, the citation count n years after publication is used for training with papers that were published more than n years after publication. On the other hand, for papers published less than n years after publication, the citation counts as of the publication of the target papers are used for training.

3.3 Target Citation Counts for Prediction

In this study, to determine an appropriate value of n for predicting citation counts, we first investigated the datasets described in Section 2. Specifically, we assumed that the citation counts five years after publication are stable, and we extracted papers published more than five years after publication from each dataset. We then calculated Spearman's rank correlation between the citation counts m months after publication and five years after publication.[6] As a result, we found that Spearman's rank correlation between the citation count one year after publication against the count five years after publication was 0.86 for the CL dataset and 0.71 for the Bio dataset. This indicates that the citation count one year after publication is a good indicator of a paper's final citation count. Hence, we adopt the citation count one year after publication as the target citation count for prediction.

4 Citation Count Complementation

We propose a method to estimate the citation count one year after the publication of papers that were published less than one year ago. Our method uses the citation counts of those papers at the time the target paper was published to estimate the counts one year after they were published. Specifically, we estimate the citation counts of a paper m months after publication with a citation count c_m by the following two methods:

Case-based Extract all papers in the training set that have a citation counts c_m at m months after publication, and use the median of those papers' counts one year after publication as the estimate.

Ratio-based For the training set, calculate the ratio of the average citation count m months after publication to the average count one year after publication, and multiply it by the citation count c_m to obtain the estimate.

While case-based estimation is expected to be accurate for less-cited papers, where there are many other papers with the same citation count, it is not suitable for highly-cited papers that have no or few other papers with the same citation count. Thus, if the citation count c_m is associated with a paper in the list of top 10% papers, it is estimated using the ratio-based method. Otherwise, it is estimated using the case-based method. The rank order of c_m is calculated from the distribution of citation counts m months after publication for the papers in the training set.

To confirm the appropriateness of this citation count complementation, we calculated Spearman's rank correlation between the correct citation counts one year after publication against the predicted citation count before and after complementation (c_m and complemented citation count). For this investigation, we used the training portion of the 12 subsets to train and evaluate the model on the CL dataset, and we compared the average Spearman's rank correlations for each subset. As a result, we found that the correlation improved from 0.88 to 0.92, which demonstrates that the citation count complementation is appropriate.

5 Degree of Early Adoption

In realistic citation count prediction, the full citation counts of papers published less than one year after publication cannot be used for training, yet papers that are frequently cited in such a short term are likely to be impactful. In addition, papers that

[5] Details of the experiment are provided in Appendix B.

[6] Detailed results are provided in Appendix C.

	Top 0–1%	1–2.5%	2.5–5%	5–10%	10–25%	25–100%	No citation
Within 3 months	15.5 (4.6%)	14.3 (3.8%)	10.6 (3.6%)	8.8 (5.0%)	7.5 (6.6%)	6.5 (4.9%)	5.0 (71.5%)
Within 6 months	14.3 (9.6%)	12.6 (7.3%)	9.8 (6.2%)	7.6 (8.8%)	6.3 (10.7%)	4.6 (9.4%)	3.7 (48.1%)
Within 9 months	13.8 (15.4%)	11.2 (10.3%)	7.7 (7.8%)	6.6 (10.1%)	5.3 (12.9%)	3.5 (12.4%)	2.5 (31.0%)
Within 12 months	12.7 (21.6%)	10.1 (12.0%)	6.4 (9.1%)	5.7 (10.5%)	4.4 (13.5%)	2.5 (12.9%)	2.0 (20.6%)

Table 1: Average citation counts one year after publication for papers citing at least one paper with the top $k_1\%$ to $k_2\%$ citation counts published within m months in the CL dataset. "No citation" indicates papers that did not cite any paper published within m months. The numbers in parentheses give the ratio of papers belonging to each group in each column.

cite such frequently cited papers earlier—i.e., papers with a high degree of early adoption—can be considered as adequately recognizing the latest trends and are likely to receive more attention in the future because of their novelty and technical contributions. To validate this hypothesis, we investigated whether papers that cite frequently cited papers at an early date tend to be cited more in the future.

Specifically, we examined the average citation count one year after publication for those papers that cite at least one paper with the top $k_1\%$ to $k_2\%$ citation counts published within m months. For this investigation, we used 15,962 papers published in arXiv's cs.CL category between June 2015 and May 2020, which form the training portion of the subset described in Section 2. In the case of multiple citations of papers published within m months, we used the highest rank order of the citation counts among them. For (k_1, k_2), we used 6 pairs: (0, 1), (1, 2.5), (2.5, 5), (5, 10), (10, 25), and (25, 100). For m, we used four values: 3, 6, 9, and 12. We then calculated the average citation count for each combination of (k_1, k_2) and m.

Table 1 lists the results. In the table, "no citation" indicates papers that did not cite any paper published within m months. We confirmed an overall trend that papers citing more recent papers and papers citing more frequently-cited papers have higher average citation counts. The average citation count of papers that cited papers in the top 1% of citations within 3 months of publication was 15.5, which was about 2.4 times higher than the average citation count of 6.5 for all papers. On the basis of these results, we attempted to leverage the degree of early adoption in citation count prediction, and we describe the specific methods for this in Section 6.1.

6 Experiments

We conducted experiments on the datasets described in Section 2 to validate the effectiveness of using citation count complementation and the degree of early adoption in realistic citation count prediction.

6.1 Setup

Task Following Maillette de Buy Wenniger et al. (2020), we defined the citation score as $\log(c_n+1)$, where c_n is the citation count n years after a paper's publication. In this study, we sought to predict the citation score one year after the publication by using the target paper's title and abstract.

Prediction Model We adopted a model based on BERT (Devlin et al., 2019) to predict the citation scores. We treated the paper's title as the first sentence of the input and the abstract as the second sentence. For the output of BERT, we used the vector representation of a special token `[CLS]`. The `[CLS]` vector was then passed through a fully connected layer and linearly transformed to obtain a prediction of the citation score. During training, we applied dropout (Srivastava et al., 2014) to the `[CLS]` vector and minimized the mean squared error (MSE) between the predicted and actual citation scores.

We also represented the degree of early adoption via a special token sequence, which was inserted at the beginning of the input sentence to BERT. Specifically, we created seven special tokens: "top 0–1%," "top 1–2.5%," "top 2.5–5%," "top 5–10%," "top 10–25%," "top 25–100%," and "no citation." This enabled us to represent the degree of early adoption by arranging the four special tokens corresponding to the highest-ranking citation counts of the papers cited by the target paper within 3, 6, 9, and 12 months, respectively. For example, if a paper cited no paper published within 3 months, a paper published within 6 to 9 months with a top 5–

10% citation count, and a paper published within 12 months with a top 0–1% citation count, the special token sequence would be "`[no citation][top 5-10%][top 5-10%][top 0-1%]`."

Experimental Setting We used two BERT-based pre-trained language models (PLMs): BERT[7] pre-trained on a general-domain corpus such as Wikipedia, and SciBERT[8] pre-trained on a scientific-domain corpus built from a large number of papers. All models were trained with 3 epochs, a batch size of 32, the AdamW optimizer (Loshchilov and Hutter, 2019), and a learning-rate schedule with warm-up at 10% of the total training steps and linear decays in the remaining steps. Following Devlin et al. (2019), the learning rate was set to 2e-5, which achieved the highest Spearman's rank correlation for all models on the development set, after searches conducted at rates of 2e-5, 3e-5, and 5e-5. We experimented with three different random seeds for each model and calculated the mean and standard deviation of the evaluation scores.[9]

Compared Methods We compared the following five methods to validate the effectiveness of using citation count complementation and leveraging the degree of early adoption.

- **Baseline:** A method that used only papers more than one year after publication for training.
- **+CCC:** A method that used all papers in the training set, including those published less than one year after publication, with **C**itation **C**ount **C**omplementation.
- **+CCC*:** A method that used the same number of papers as the Baseline model, in order from the newest in the training set, with **C**itation **C**ount **C**omplementation.
- **+DEA:** A method that was based on the Baseline model but used the **D**egree of **E**arly **A**doption.
- **+CCC+DEA:** A method that used all papers in the training set with **C**itation **C**ount **C**omplementation and the **D**egree of **E**arly **A**doption.

We also considered applying the proposed method to the existing citation count prediction models based on deep learning such as NNCP (Abrishami and Aliakbary, 2019), BIL_A (Ma et al., 2021), and SChuBERT (van Dongen et al., 2020), but discarded the idea for the following reasons. First, NNCP and BIL_A were designed under the assumption that citation counts several years after a target paper's publication are available, and thus these models were not applicable to our setting. SChuBERT was excluded from the experiments because preliminary experiments showed that its performance was equal to or lower than the Baseline, even though it is a model that predicts citation counts using the entire body of a paper. The low performance of SChuBERT is probably due to the fact that it does not perform fine-tuning since it would be computationally expensive to perform fine-tuning for SChuBERT.

Evaluation We evaluated the models with three metrics: Spearman's rank correlation (ρ) to assess the overall ranking quality, the mean squared error (MSE) to assess the amount of error, and a metric defined as the percentage of the actual top n% of papers in the top k% of the output (n%@k%) to intuitively understand the results.

As mentioned in Section 2, because the average number of papers for evaluation in each subset of the datasets was not large, the evaluation scores would not have been stable if each subset were evaluated individually. Therefore, to yield stable results, we computed each metric across all subsets of the papers. That is, while each subset was used to train the prediction model and the citation counts of the papers for evaluation were predicted by using the model for each subset, the evaluation scores were calculated by combining the predictions for all 12 subsets.

6.2 Experimental Results

Table 2 summarizes the experimental results. For both the CL and Bio datasets, the models based on BERT and SciBERT improved the citation count prediction performance by leveraging either the citation count complementation or the degree of early adoption. The performance was further improved by using both. The SciBERT-based model outperformed the BERT-based model, which demonstrated the effectiveness of pre-training on a scientific-domain corpus for citation count prediction.[10]

By comparing the Baseline and +CCC* models,

[7] https://huggingface.co/bert-base-uncased

[8] https://huggingface.co/allenai/scibert_scivocab_uncased

[9] Training took about 10 minutes per epoch and inference took a few seconds per evaluation set on a single GV100 GPU.

[10] We also experimented with domain-specific models such as PubMedBERT (Gu et al., 2021) on the Bio dataset, but we could not confirm further performance improvement.

Dataset	PLM	Method	ρ	MSE	5%@5%	5%@25%	10%@10%	10%@50%
CL	BERT	Baseline	$36.6_{\pm 0.4}$	$1.504_{\pm 0.022}$	$21.1_{\pm 1.7}$	$63.7_{\pm 2.3}$	$28.1_{\pm 0.9}$	$83.2_{\pm 0.4}$
		+CCC	$39.1_{\pm 0.2}$	$1.275_{\pm 0.018}$	$\mathbf{28.8}_{\pm 1.9}$	$72.6_{\pm 1.0}$	$34.5_{\pm 0.4}$	$84.6_{\pm 0.6}$
		+CCC*	$39.6_{\pm 0.1}$	$1.176_{\pm 0.041}$	$28.2_{\pm 1.5}$	$73.4_{\pm 1.0}$	$34.4_{\pm 0.7}$	$84.9_{\pm 1.1}$
		+DEA	$40.4_{\pm 0.5}$	$1.394_{\pm 0.019}$	$22.4_{\pm 0.6}$	$69.4_{\pm 0.9}$	$31.1_{\pm 0.8}$	$86.7_{\pm 0.7}$
		+CCC+DEA	$\mathbf{41.8}_{\pm 0.3}$	$\mathbf{1.173}_{\pm 0.008}$	$28.6_{\pm 2.1}$	$\mathbf{75.3}_{\pm 2.2}$	$\mathbf{35.7}_{\pm 1.1}$	$\mathbf{87.0}_{\pm 1.0}$
	SciBERT	Baseline	$38.3_{\pm 0.3}$	$1.390_{\pm 0.042}$	$27.4_{\pm 1.2}$	$67.5_{\pm 1.5}$	$32.0_{\pm 1.0}$	$84.7_{\pm 0.7}$
		+CCC	$40.1_{\pm 0.5}$	$1.147_{\pm 0.010}$	$33.2_{\pm 1.7}$	$72.8_{\pm 0.6}$	$37.7_{\pm 0.4}$	$86.2_{\pm 0.2}$
		+CCC*	$40.9_{\pm 0.1}$	$\mathbf{1.063}_{\pm 0.013}$	$33.1_{\pm 0.9}$	$75.5_{\pm 0.9}$	$\mathbf{37.8}_{\pm 0.9}$	$86.0_{\pm 0.4}$
		+DEA	$41.1_{\pm 0.4}$	$1.307_{\pm 0.015}$	$28.0_{\pm 1.4}$	$70.3_{\pm 0.4}$	$33.8_{\pm 1.2}$	$86.2_{\pm 0.2}$
		+CCC+DEA	$\mathbf{42.8}_{\pm 0.1}$	$1.104_{\pm 0.012}$	$\mathbf{34.2}_{\pm 0.5}$	$\mathbf{76.0}_{\pm 1.1}$	$36.7_{\pm 1.1}$	$\mathbf{87.9}_{\pm 0.2}$
Bio	BERT	Baseline	$24.1_{\pm 2.0}$	$0.593_{\pm 0.012}$	$20.1_{\pm 1.1}$	$41.3_{\pm 4.6}$	$26.8_{\pm 2.4}$	$67.5_{\pm 3.4}$
		+CCC	$36.4_{\pm 1.1}$	$0.487_{\pm 0.010}$	$\mathbf{50.4}_{\pm 1.0}$	$83.1_{\pm 0.0}$	$48.4_{\pm 0.9}$	$86.7_{\pm 0.9}$
		+CCC*	$32.9_{\pm 0.9}$	$0.499_{\pm 0.011}$	$49.8_{\pm 1.0}$	$80.5_{\pm 1.3}$	$47.1_{\pm 0.6}$	$84.2_{\pm 2.1}$
		+DEA	$32.7_{\pm 3.0}$	$0.559_{\pm 0.018}$	$21.9_{\pm 0.4}$	$47.4_{\pm 4.7}$	$29.9_{\pm 1.6}$	$77.1_{\pm 6.7}$
		+CCC+DEA	$\mathbf{40.6}_{\pm 0.6}$	$\mathbf{0.461}_{\pm 0.005}$	$50.0_{\pm 0.0}$	$\mathbf{87.7}_{\pm 0.6}$	$\mathbf{49.6}_{\pm 0.7}$	$\mathbf{89.9}_{\pm 1.5}$
	SciBERT	Baseline	$30.3_{\pm 1.0}$	$0.588_{\pm 0.011}$	$21.0_{\pm 2.6}$	$51.7_{\pm 2.6}$	$29.9_{\pm 0.9}$	$73.2_{\pm 2.6}$
		+CCC	$40.5_{\pm 0.3}$	$0.446_{\pm 0.007}$	$\mathbf{54.3}_{\pm 0.4}$	$86.8_{\pm 0.7}$	$52.5_{\pm 0.7}$	$89.4_{\pm 0.7}$
		+CCC*	$37.2_{\pm 0.7}$	$0.472_{\pm 0.006}$	$53.7_{\pm 0.4}$	$84.4_{\pm 1.3}$	$48.4_{\pm 0.3}$	$88.3_{\pm 1.7}$
		+DEA	$37.0_{\pm 2.3}$	$0.555_{\pm 0.018}$	$25.1_{\pm 1.5}$	$57.8_{\pm 6.2}$	$33.7_{\pm 1.8}$	$79.4_{\pm 5.3}$
		+CCC+DEA	$\mathbf{42.5}_{\pm 1.2}$	$\mathbf{0.436}_{\pm 0.010}$	$52.4_{\pm 0.4}$	$\mathbf{90.3}_{\pm 2.6}$	$\mathbf{52.6}_{\pm 0.6}$	$\mathbf{91.8}_{\pm 1.8}$

Table 2: Experimental results from comparing methods that use papers published less than one year after publication in realistic citation count prediction. Each score besides the MSE is multiplied by 100.

BERT for Coreference Resolution: Baselines and Analysis

Abstract: We apply BERT to coreference resolution, achieving strong improvements on the OntoNotes (+3.9 F1) and GAP (+11.5 F1) benchmarks. A qualitative analysis of model predictions indicates that, compared to ELMo and BERT-base, BERT-large is particularly better at distinguishing between related but distinct entities (e.g., President and CEO). However, there is still room for improvement in modeling document-level context, conversations, and mention paraphrasing. Our code and models are publicly available.

Ground truth:	top 0.9%
Baseline:	top 14.5%
+CCC:	top 2.8%
+DEA:	top 7.1%
+CCC+DEA:	top 0.8%

Figure 2: Example of a paper for which the citation count complementation and degree of early adoption improved the prediction. The left part shows the papers title and abstract (Joshi et al., 2019), and the right part shows the relative position of the citation count one year after publication of the target paper (ground truth) and the relative positions predicted by SciBERT-based models.

which used the same number of papers for training, we can see that the +CCC* model performed better on both datasets; thus, we confirmed the effectiveness of using papers published less than one year after publication with citation count complementation for training. We had predicted that the +CCC model, which used a larger number of papers for training, would perform better than the +CCC* model. This was true for the Bio dataset, but surprisingly for the CL dataset, the +CCC* model performed better. We speculate that older papers could serve as noise if the number of papers is sufficiently large, but we leave further investigation of this point to a future work. From the result for the Baseline, +CCC, and +CCC* models on the Bio dataset, we confirmed performance improvement due to the increased number of papers for training and the leverage of newer papers. In particular, the performance gains from using new papers for training were considerable.

As for the actual predictive performance, the SciBERT-based model using both citation count complementation and the degree of early adoption achieved a score of 90.3 for the 5%@25% metric on the Bio dataset. This means that if we read only the top 25% of the papers predicted by the model for a given set of papers, we could cover 90.3% of the papers expected to have future citation counts within the top 5%. Hence, we believe that this method is highly useful from a practical viewpoint.

Figure 2 shows an example of a paper for which the citation count complementation and degree of early adoption improved the prediction. Although the citation count one year after the paper's publi-

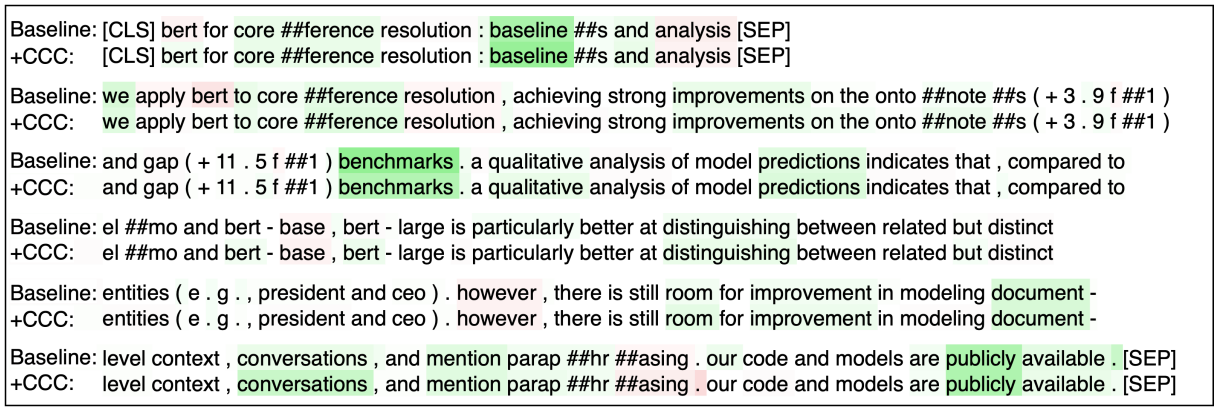

Figure 3: Visualization of the contribution of each token in predicting the citation count of the paper shown in Figure 2. Darker green represents a higher contribution to the prediction, while darker red represents a lower contribution.

cation was in the top 0.9% in the evaluation set, the Baseline model underestimate the citation count. This is likely because the paper was published 10 months after the original paper on BERT, and the Baseline model thus could not leverage the "latest" information that BERT was going to get enormous attention. The prediction was improved by applying either of the two proposed methods, and it was quite accurate when both methods were applied. The use of papers published less than one year after publication for training by citation count complementation would enable the model to use information about BERT for prediction. In addition, this paper cited the top 5% to 10% of papers within 3 months of publication and the top 0% to 1% of papers within 6 months of publication, which indicates that it captured the latest trends. We believe that the proposed method successfully incorporated these properties of the paper into citation count prediction by leveraging the degree of early adoption.

6.3 Analysis and Discussion

To investigate what words the model came to emphasize by leveraging papers shortly after publication for training, we performed an analysis using Integrated Gradients (Sundararajan et al., 2017). The Integrated Gradients method computes each input feature's contribution to a deep network's prediction by integrating gradients; thus, it enables analysis of each input token's contribution to a prediction by BERT. Similar to Schwarzenberg et al. (2021) and Bharadwaj and Shevade (2022), we used a sequence of `[PAD]` tokens as the baseline input for Integrated Gradients to estimate the contribution of each token.

Figure 3 shows a visualization of the contribution of each token in predicting the citation count of the example paper shown in Figure 2, for the Baseline model, which does not use papers published after the BERT paper for training, and the +CCC model, which uses papers published after the BERT paper for training. The darker green represents a higher contribution to the prediction, while the darker red represents a lower contribution. The figure shows that the Baseline model did not know about BERT, and the token *bert* had a negative impact, whereas the +CCC model knew that BERT was a state-of-the-art model, and the token had a positive impact. We also observed that both models emphasized tokens that are intuitively important, such as the higher contribution of *publicly available*, which is thought to facilitate subsequent research and growth in citation counts when codes and models are made publicly available.

Furthermore, we quantitatively analyzed the tokens whose contribution to the prediction was increased by using papers shortly after publication for training. To extract these tokens, we calculated each token's contribution in the +CCC model and its contribution in the Baseline model for the same paper. Then, we took the difference to obtain the score increase due to the use of papers published less than one year after publication. We computed this increase by using all the papers for evaluation in each of the two datasets, took the average for each token, and extracted the top 10 tokens for that average. If a word was divided into subwords, its contribution was determined by summing the subwords' contributions. In addition, stop words, tokens containing symbols, and tokens with a document frequency of less than 10 were deleted.

Table 3 lists the extracted words. In the CL

Rank	CL	Bio
1	trec	coronavirus
2	coronavirus	coronaviruses
3	revisiting	sars
4	semeval	cov
5	rethinking	computationally
6	finnish	tumors
7	wmt	nucleocapsid
8	bert	hydroxychloroquine
9	propaganda	cannabis
10	specaugment	pandemic

Table 3: Tokens that the model came to emphasize by using papers shortly after publication for training by citation count complementation.

dataset, the conference names *trec*, *semeval*, and *wmt* were at the top of the list. This could mean that more and more papers have evaluated models on datasets that were published at those conferences in recent years. Other words such as *revisiting* and *rethinking* may be associated with an increase in the number of papers that have revised existing models and methods in recent years. In fact, the number of papers published at ACL that included these words in their titles increased from three (0.15%) in 2013-2018 to 15 (0.53%) in 2019-2022. The model also increasingly focused on technologies that have gained attention in recent years, such as *bert* and *specaugment*. In particular, SpecAugment (Park et al., 2019) is a high-profile technology in the speech-processing field that has been cited more than 2,000 times since it was published in April 2019, and the model was able to capture it here as an important technology.

As for the Bio dataset, a number of COVID-19-related words appeared at the top of the list. This indicates that the model captured the increasing number of relevant papers and increasing overall citation counts due to the COVID-19 pandemic. Also, we attribute the large performance improvement with citation count completion on the Bio dataset to the capability to focus more on COVID-19-related words.

7 Related Work

Early works on citation count prediction formulated the task and explored effective features. Castillo et al. (2007) formulated citation count prediction as a regression problem and used author reputation to predict the citation count. Fu and Aliferis (2008) formulated citation count prediction as a classification problem and investigated several features that are effective for such prediction, including a paper's title, abstract, and author information.

Other studies have sought to improve the prediction performance by using various features. One such feature is a citation graph constructed from citation relationships among papers. Davletov et al. (2014) proposed a method to use the graph's temporal and topological features. Pobiedina and Ichise (2015) achieved high prediction performance by mining frequent graph patterns. Singh et al. (2015) proposed a method to use the citation context, which is the text in a paper that mentions other cited papers. Bhat et al. (2015) found that the interdisciplinarity of authors is effective in predicting citation counts. Li et al. (2019) proposed a method to use peer-reviewed text from multiple aspects.

Several studies have focused on aspects other than features. Chakraborty et al. (2014) and te Li et al. (2015) found several patterns in the growth of citation counts by analyzing a large number of papers, and they proposed a two-step prediction method, first classifying papers into each pattern and then predicting counts for each pattern. Xiao et al. (2016) proposed a method to predict the citation count at an arbitrary point in time from the publication of a paper, with the aim of predicting its future potential impact.

In recent years, there has been research on the use of deep learning techniques to predict citation counts. Abrishami and Aliakbary (2019) proposed an RNN-based method to predict a paper's future citation count by using the citation counts for each elapsed year since its publication. van Dongen et al. (2020) proposed a method to predict the citation count by dividing a paper's text into chunks and encoding the paper's entire body with BERT. Ma et al. (2021) proposed a method to predict the citation count by extracting semantic features from a paper's title and abstract via Doc2Vec and Bi-LSTM with an attention mechanism.

8 Conclusion

In this paper, we introduced a realistic citation count prediction task that is applicable to newly published papers, by using only citation count information that is strictly available at the time of pub-

lication of a target paper for training. We further proposed two methods to use papers published less than one year after publication for citation count prediction, as these papers cannot be directly used for training because their citation counts one year after publication are unknown. The first method is citation count complementation, which uses recent papers for training by estimating their citation counts one year after publication. The second method is to leverage the degree of early adoption, which incorporates the tendency for papers that cite highly cited papers earlier to have higher average citation counts. Through experiments using papers collected from arXiv and bioRxiv, we demonstrated that the use of papers published less than one year after publication improves the performance of realistic citation count prediction. For future work, we intend to build models that incorporate information from papers that was not used in this study, such as the body, figures, tables, and author information.

Limitations

Both methods proposed in this paper focus on fields in which technology is rapidly evolving and the latest research results are increasingly important. Because of this, these methods' effectiveness could be limited in fields for which the latest research results are not particularly important. Also, the model in this study only uses the titles and abstracts of papers as inputs, and it does not leverage the body, figures, or tables.

Acknowledgements

This work was partly supported by JST Moonshot R&D (Grant Number JPMJMS2033).

References

Ali Abrishami and Sadegh Aliakbary. 2019. Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, 13(2):485–499.

Dagfinn W. Aksnes. 2006. Citation rates and perceptions of scientific contribution. *J. Assoc. Inf. Sci. Technol.*, 57:169–185.

Shikhar Bharadwaj and Shirish Shevade. 2022. Efficient constituency tree based encoding for natural language to bash translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2022)*, pages 3159–3168.

Harish S. Bhat, Li-Hsuan Huang, Sebastian Rodriguez, Rick Dale, and Evan Heit. 2015. Citation prediction using diverse features. *2015 IEEE International Conference on Data Mining Workshop (ICDMW 2015)*, pages 589–596.

Carlos Castillo, Debora Donato, and Aristides Gionis. 2007. Estimating number of citations using author reputation. In *String Processing and Information Retrieval (SPIRE 2007)*, pages 107–117.

Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2014. Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 351–360.

Daryl E. Chubin and Eugene Garfield. 1979. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 2:91–94.

Feruz Davletov, Ali Selman Aydin, and Ali Cakmak. 2014. High impact academic paper prediction using temporal and topological features. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM 2014)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL 2019)*, pages 4171–4186.

Lawrence D. Fu and Constantin F. Aliferis. 2008. Models for predicting and explaining citation count of biomedical articles. *AMIA ... Annual Symposium proceedings. AMIA Symposium vol. 2008 (AMIA 2008)*, pages 222–226.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 5803–5808.

Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji-Rong Wen. 2019. A neural citation count prediction model based on peer review text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 4914–4924.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR 2019)*.

Anqi Ma, Yu Liu, Xiujuan Xu, and Tao Dong. 2021. A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics*, 126:6803–6823.

Gideon Maillette de Buy Wenniger, Thomas van Dongen, Eleri Aedmaa, Herbert Teun Kruitbosch, Edwin A. Valentijn, and Lambert Schomaker. 2020. Structure-tags improve text classification for scholarly document quality prediction. In *Proceedings of the First Workshop on Scholarly Document Processing (SDP 2020)*, pages 158–167.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech 2019*, pages 2613–2617.

Nataliia Pobiedina and Ryutaro Ichise. 2015. Citation count prediction as a link prediction problem. *Applied Intelligence*, 44:252–268.

Robert Schwarzenberg, Nils Feldhus, and Sebastian Möller. 2021. Efficient explanations from empirical explainers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP 2021)*, pages 240–249.

Mayank Singh, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. 2015. The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM 2015)*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR 2014)*, 15(56):1929–1958.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML 2017)*, page 3319–3328.

Cheng te Li, Yu-Jen Lin, Rui Yan, and Mi-Yen Yeh. 2015. Trend-based citation count prediction for research articles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2015)*.

Thomas van Dongen, Gideon Maillette de Buy Wenniger, and Lambert Schomaker. 2020. SChuBERT: Scholarly document chunks with BERT-encoding boost citation count prediction. In *Proceedings of the First Workshop on Scholarly Document Processing (SDP 2020)*, pages 148–157.

Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M. Chu, and Hongyuan Zha. 2016. On modeling and predicting individual paper citation count over time. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*.

A Detailed Dataset Statistics

Table 4 lists the numbers of papers for training and evaluation for each subset in the CL and Bio datasets described in Section 2.

Dataset	Subset	Training	Evaluation
CL	6/2019	10,459	620
	7/2019	11,026	404
	8/2019	11,404	479
	9/2019	11,854	720
	10/2019	12,529	550
	11/2019	13,031	564
	12/2019	13,552	345
	1/2020	13,820	260
	2/2020	14,049	326
	3/2020	14,339	334
	4/2020	14,617	747
	5/2020	15,305	713
	6/2020	15,962	440
Bio	5/2020	4,451	292
	6/2020	4,743	303
	7/2020	5,046	286
	8/2020	5,331	268
	9/2020	5,597	233
	10/2020	5,827	261
	11/2020	6,088	221
	12/2020	6,307	219
	1/2021	6,524	245
	2/2021	6,769	246
	3/2021	7,012	258
	4/2021	7,264	252

Table 4: Numbers of papers for training and evaluation for each subset in the CL and Bio datasets. The subset names correspond to the year and month of publication of the papers that a subset used for evaluation.

B Details of Leakage Investigation in Existing Settings

To investigate the impact of leakage in the existing setting on the performance of citation count prediction, we conducted an experiment using the CL dataset described in Section 2. The experiment basically used the Baseline model described

PLM	Setting	Avg. train size	ρ	MSE	5%@5%	5%@25%	10%@10%	10%@50%
BERT	w/ future citation count	13,277	$40.5_{\pm 0.3}$	$1.373_{\pm 0.010}$	$28.7_{\pm 2.0}$	$72.8_{\pm 1.0}$	$34.6_{\pm 0.1}$	$87.1_{\pm 0.3}$
	w/ future citation count	8,571	$39.0_{\pm 0.2}$	$1.358_{\pm 0.030}$	$26.0_{\pm 0.2}$	$73.5_{\pm 1.0}$	$33.7_{\pm 1.2}$	$85.1_{\pm 0.4}$
	w/o future citation count	8,571	$36.6_{\pm 0.4}$	$1.504_{\pm 0.022}$	$21.1_{\pm 1.7}$	$63.7_{\pm 2.3}$	$28.1_{\pm 0.9}$	$83.2_{\pm 0.4}$
SciBERT	w/ future citation count	13,277	$41.8_{\pm 0.3}$	$1.220_{\pm 0.024}$	$31.1_{\pm 1.1}$	$73.5_{\pm 1.5}$	$37.1_{\pm 0.8}$	$87.9_{\pm 0.4}$
	w/ future citation count	8,571	$40.4_{\pm 0.9}$	$1.232_{\pm 0.019}$	$31.3_{\pm 2.1}$	$72.6_{\pm 1.0}$	$35.8_{\pm 0.7}$	$86.2_{\pm 1.1}$
	w/o future citation count	8,571	$38.3_{\pm 0.3}$	$1.390_{\pm 0.042}$	$27.4_{\pm 1.2}$	$67.5_{\pm 1.5}$	$32.0_{\pm 1.0}$	$84.7_{\pm 0.7}$

Table 5: Experimental results of the investigation of the leaks in the existing setting (w/ future citation count). Each score besides the MSE is multiplied by 100.

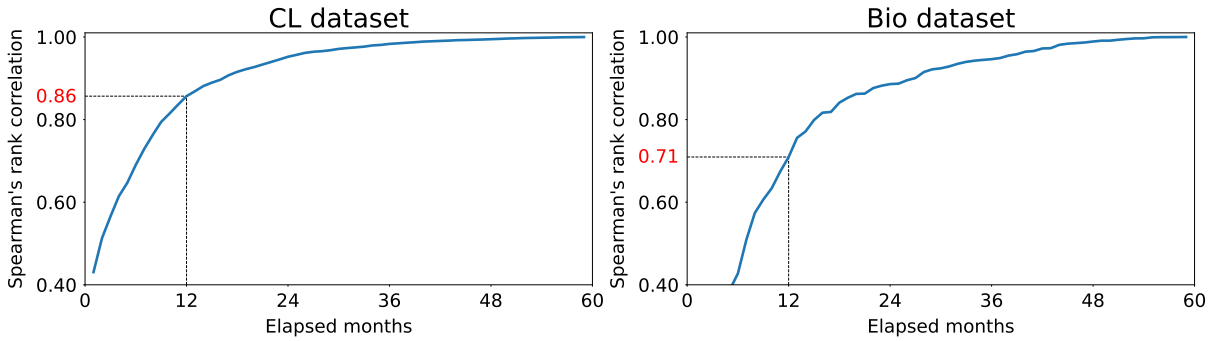

Figure 4: Spearman's rank correlation between the citation counts m months after publication against the citation count five years after publication. The left part shows the results on the CL dataset and the right part shows the results on the Bio dataset.

in Section 6.1, and only the papers for training were changed. We compared settings that used future citation counts with those that do not. In the setting that did not use future citation counts (*w/o future citation count*), only papers published more than one year after publication as of the target paper's publication were used for training. For example, if the subset that used papers published in June 2020 for evaluation, papers published between July 2019 and May 2020 were excluded from the training set, and only papers published between June 2015 and June 2019 were used for training. This reduced the average number of papers for training from 13,227 to 8,571. In the setting that used future citation counts (*w/ future citation count*), we used the citation counts one year after publication for all papers in the training set, including papers published less than one year after publication as of the target paper's publication.

In the *w/ future citation count* setting, the number of papers that can be used for training was larger than in the *w/o future citation count* setting, and thus the impact of the leakage could not be fairly investigated. For a fair comparison, we also experimented with settings that align the number of papers for training used in the *w/ future citation count* setting with the *w/o future citation count* setting. The number of papers for training was aligned by grouping the papers for training by year and month of publication and randomly reducing the papers in each group by the same ratio. By aligning the number of papers, we could fairly compare *w/* and *w/o future citation count* settings.

Table 5 shows the experimental results. For all metrics, the *w/ future citation count* setting, which was trained using all citation count that was actually unavailable, outperforms the *w/o future citation count* setting, which was trained using only available information. The results show that the existing setting improperly improves the performance of the prediction model. In particular, even when the number of papers for training was aligned, the *w/ future citation count* setting outperformed the *w/o future citation count* setting. This demonstrates that the future citation count of papers published close to the target causes leakage of research trends that grow in citation count in the future.

C Transition of Spearman's Rank Correlation

Figure 4 shows Spearman's rank correlation between the citation counts m months after publication and five years after publication in the CL and Bio datasets.