What Have We Learned from OpenReview?

Gang Wang[1], Qi Peng[1], Yanfeng Zhang[2], and Mingyang Zhang[3]

Northeastern University, China
1910636@stu.neu.edu.cn
ffpengqi@stumail.neu.edu.cn
zhangyf@mail.neu.edu.cn
theremay@outlook.com

Abstract. Anonymous peer review is used by the great majority of computer science conferences. OpenReview is such a platform that aims to promote openness in peer review process. The paper, (meta) reviews, rebuttals, and final decisions are all released to public. We collect 5,527 submissions and their 16,853 reviews from the OpenReview platform. We also collect these submissions' citation data from Google Scholar and their non-peer-reviewed versions from arXiv.org. By acquiring deep insights into these data, we have several interesting findings that could help understand the effectiveness of the public-accessible double-blind peer review process. Our results can potentially help writing a paper, reviewing it, and deciding on its acceptance.

Keywords: Peer review · OpenReview · Opinion divergence.

1 Introduction

Peer review is a widely adopted quality control mechanism in which the value of scientific paper is assessed by several reviewers with a similar level of competence. The primary role of the review process is to decide which papers to publish and to filter information, which is particularly true for a top conference that aspires to attract a broad readership to its papers. The novelty, significance, and technical flaws are expected to be identified by the reviewers, which can help the PC chair make the final decision.

Anonymous peer review (no matter single-blind or double-blind), despite the criticisms often leveled against it, is used by the great majority of computer science conferences, where the reviewers do not identify themselves to the authors. It is understandable that some authors are uncomfortable with a system in which their identities are known to the reviewers while the latter remain anonymous. Authors may feel themselves defenseless against what they see as the arbitrary behavior of reviewers who cannot be held accountable by the authors for unfair comments. On the other hand, apparently, there would be even more problems if letting authors know their reviewers' identities. Reviewers would give more biased scores for fear of retaliation from the more powerful colleagues. Given this contradiction, opening up the reviews to public seems to be a good solution.

The openness of reviews will force reviewers to think more carefully about the scientific issues and to write more thoughtful reviews, since PC chairs know the identities of reviewers and bad reviews would affect their reputations.

OpenReview[1] is such a platform that aims to promote openness in peer review process. The paper, (meta) reviews, rebuttals, and final decisions are all released to public. Colleagues who do not serve as reviewers can judge the paper's contribution as well as judge the fairness of the reviews by themselves. Reviewers will have more pressure under public scrutiny and force themselves to give much fairer reviews. On the other hand, previous works on peer-review analysis [17,13,20,14,11,19] are often limited due to the lack of rejected paper instances and their corresponding reviews. Given these public reviews (for both accepted papers and rejected ones), studies towards multiple interesting questions related to peer-review are made available.

Given these public reviews, there are multiple interesting questions raised that could help us understand the effectiveness of the public-accessible double-blind peer review process: a) As known, AI conferences have extremely heavy review burden in 2020 due to the explosive number of submissions [2]. These AI conferences have to hire more non-experts to involve in the double-blind review process. How is the impact of these non-experts on the review process (Sec. 3.1)? b) Reviewers often evaluate a paper from multiple aspects, such as motivation, novelty, presentation, and experimental design. Which aspect has a decisive role in the review score (Sec. 3.2)? c) The OpenReview platform provides not only the submission details (e.g., title, keywords, and abstract) of accepted papers but also that of rejected submissions, which allows us to perform a more fine-grained cluster analysis. Given the fine-grained hierarchical clustering results, is there significant difference in the acceptance rate of different research fields (Sec. 3.3)? d) A posterior quantitative method for evaluating papers is to track their citation counts. A high citation count often indicates a more important, groundbreaking, or inspired work. OpenReview releases not only the submission details of accepted papers but also that of rejected submissions. The rejected submissions might be put on arXiv.org or published in other venues to still attract citations. This offers us opportunities to analyze the correlation between review scores and citation numbers. Is there a strong correlation between review score and citation number for a submission (Sec. 3.4)? e) Submissions might be posted on arXiv.org before the accept/reject notification, which might be the rejected ones from other conferences. They are special because they could be improved according to the rejected reviews and their authors are not anonymous. Are these submissions shown higher acceptance rate (Sec. 3.5)?

In this paper, we collect 5,527 (accepted and rejected) submissions and their 16,853 reviews from ICLR 2017-2020 venues[2] on the OpenReview platform as our main corpus. By acquiring deep insights into these data, we have several interesting findings and aim to answer the above raised questions quantitatively. Our submitted supplementary file also includes more data analysis results. Given

[1] https://openreview.net/

[2] International Conference on Learning Representations. https://iclr.cc/

Table 1. Statistics of ICLR reviews dataset

year	#papers	#authors	accept rate	#reviews	review len.
2017	489	1,417	50.1%	1,495	295.11
2018	939	2,882	49.0%	2,849	372.07
2019	1,541	4,332	32.5%	4,733	403.22
2020	2,558	7,765	26.5%	7,766	407.08
total	5,527	16,396	39.5%	16,843	369.37

these data analysis results, we expect to introduce more discussions on the effectiveness of peer-review process and hope that treatment will be obtained to improve the peer-review process.

2 Dataset

ICLR has used OpenReview to launch double-blind review process for 8 years (2013-2020). Similar to other major AI conferences, ICLR adopts a reviewing workflow containing double-blind review, rebuttal, and final decision process. After paper assignment, typically three reviewers evaluate a paper independently. After the rebuttal, reviewers can access the authors' responses and other peer reviews, and accordingly modify their reviews. The program chairs then write the meta-review for each paper to make the final accept/reject decision according to the three anonymous reviews. Each official review mainly contains a review score (integer between 1 and 10), a reviewer confidence level (integer between 1 and 5), and the detailed review comments. The official reviews and meta-reviews are all open to the public on the OpenReview platform. Public colleagues can also post their reviews on OpenReview. In the following, we present the dataset of submissions and reviews we collect from the OpenReview platform, these submissions' citation data from Google Scholar, and their non-peer-reviewed versions from arXiv.org[3].

Submissions and Reviews. We have collected 5,527 submissions and 16,853 official reviews from ICLR 2017-2020 venues on the OpenReview platform. We only use the review data since 2017 because the submissions before 2017 is too few. Though a double-blind review process is exploited, the authors' identities of the rejected submissions are also released after decision notification. Thus, we can also access the identity information for each rejected submission, which is critical in most of our analysis. Some statistics of the reviews data are listed in Table 1, in which **review len.** indicates the average length of reviews.

Citations. In order to investigate the correlation between review scores and citation numbers, we also collect the citation information from Google Scholar for all the 1,183 accepted papers from 2017 to 2019. Since the rejected submissions might be put on arXiv.org or published in other venues, they might also attract citations. We also collect the citation information for 2,054 rejected submissions that have been published elsewhere (210 for 2017, 324 for 2018, 493 for 2019,

[3] These datasets and the source code for the analysis experiment are available at https://github.com/Seafoodair/Openreview/

Table 2. Statistics of different confidence level reviews

	2017-2019					2020			
	level1	**level2**	**level3**	**level4**	**level5**	**level1**	**level2**	**level3**	**level4**
#reviews	74	455	2,330	4,612	1,600	1,104	2,554	2,659	1,449
fraction	0.80%	5.01%	25.67%	50.81%	17.71%	14.22%	32.89%	34.24%	18.66%

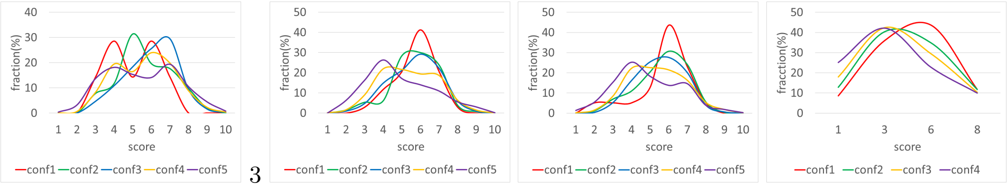
and totally 1027 rejected papers). All the citation numbers are gathered up to 20 Jan. 2020. We do not collect citation information of ICLR 2020 papers because they have not yet accumulated enough citations yet.

arXiv Submissions. In order to investigate whether the submissions that have been posted on arXiv.org before notification have a higher acceptance rate, we also crawl the arXiv versions of ICLR 2017-2020 submissions if they exist. We record the details of an arXiv preprint if its title matches an ICLR submission title. Note that, their contents might be slightly different. We totally find 1,158 matched arXiv papers and 948 among them were posted before notification (178/150 for 2017, 103/79 for 2018, 420/303 for 2019, and 457/416 for 2020) up to 18 Feb 2020.

3 Results Learned from Open Reviews

3.1 How is the Impact of Non-Expert Reviewers?

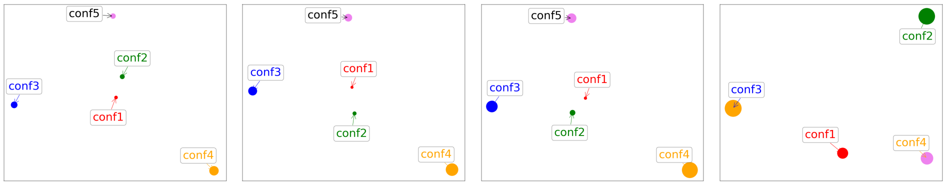
Due to the extensively increasing amount of submissions, ICLR 2020 hired much more reviewer volunteers. There were complaints about the quality of reviews (47% of the reviewers have not published in the related areas [2]). Similar scenarios have been observed in other AI conferences, such as NIPS, CVPR, and AAAI. Many authors complain that their submissions are not well evaluated because the assigned "non-expert" reviewers lack of enough technical background and cannot understand their main contributions. How is the impact of these "non-experts" on the review process? In this subsection, we aim to answer the question through quantitative data analysis.



(a) 2017 (b) 2018 (c) 2019 (d) 2020

Fig. 1. The review score distributions of different confidence level (conf) reviews

Review Score Distribution. For ICLR 2017-2019, reviewer gives a review score (integer) from 1 to 10, and is asked to select a confidence level (integer) between 1 and 5. For ICLR 2020, reviewer gives a rating score in {1, 3, 6, 8} and should select an experience assessment score (similar to confidence score) between 1 and 4. We divide the reviews into multiple subsets according to their confidence levels. Fig. 1 shows the smoothed review score distributions for each subset of reviews. For 2018-2020, we consistently observe that the scores of reviews with confidence level 1 and 2 are likely to be higher than those reviews with confidence level 4 and 5. The trend of ICLR 2017 is not clear because it contains too few samples to be statistically significant (e.g., only 7 level-1 reviews). In 2017-2019, the lowest confidence level reviews has an average review score 5.675, while the highest confidence level reviews has an average review score 4.954. In 2020, the numbers for the lowest and highest confidence level reviews are 4.726 and 3.678, respectively. Our results show that the‘low-confidence reviewers (e.g., level 1 and 2) tend to be more tolerant because they may be not confident about their decision, while the high-confidence reviewers (e.g., level 4 and 5) tend to be more tough and rigorous because they may be confident in the identified weakness.



(a) 2017 (b) 2018 (c) 2019 (d) 2020

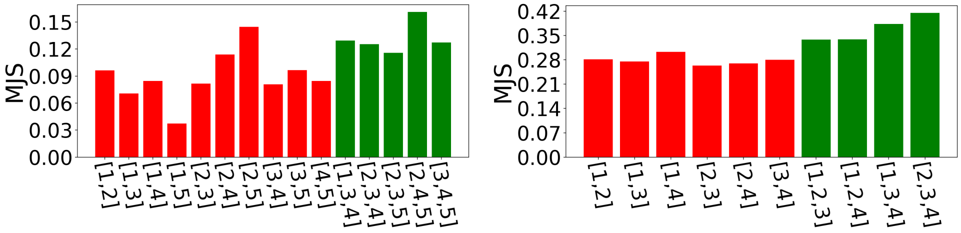
Fig. 2. The visualized layout of groups of reviews with different confidence scores. Each point indicates a group of reviews with a specific confidence level (abbrv. conf). The size of point indicates the relative number of reviews in that group. The distance between two points indicates the divergence of review scores between two groups.

Divergence Reflected by Euclidean Distance. On the other hand, peoples are worrying that non-expert reviewers are not competent to give a fair evaluation of a submission (e.g., fail to identify key contributions or fail to identify flaws) and will ruin the reputation of top conferences [2]. Particularly, these non-expert reviewers may have different opinions with the expert reviewers regarding the same paper. Actually, opinion divergence commonly exists between reviewers in the peer-review process. Each paper is typically assigned to 3 reviewers. These 3 reviewers may have significantly different review scores. In order to illustrate the difference between the reviews with different confidence scores, we first compute the euclidean distance $DIS(l_i, l_j)$ between between group l_i and group l_j as follows. Let R_{l_i,l_j} be the set of paper IDs, where each paper concurrently has both confidence-l_i review(s) and confidence-l_j review(s). Let $\overline{s_p^i}$ be paper p's average review score from l_i-confidence reviews. Then, the distance between the group of confidence-l_i reviews and that of confidence-l_j reviews is:

$$DIS(l_i, l_j) = \sqrt{\sum_{p \in R_{l_i,l_j}} \left(\overline{s_p^i} - \overline{s_p^j}\right)^2}. \qquad (1)$$

After computing the distance betwenn each pair of groups, we can construct a distance matrix. According to the distance matrix, we use t-SNE [15] to plot the visualized layout of different groups of reviews of each year in Fig. 2. For ICLR

2017-2019, we can see similar layout, where conf1 reviews are close to conf2 reviews in the central part and the groups of conf3, conf4, and conf5 locate around. The group of conf4 reviews is far apart from the most professional reviews (conf5). In ICLR 2020, there are 4 confidence levels. Surprisingly, we observe that the most professional reviews (conf4) and least professional reviews (conf1) are closest to each other. Conf2 reviews and conf3 reviews are both far apart from conf4 reviews.



(a) 2017-2019

(b) 2020

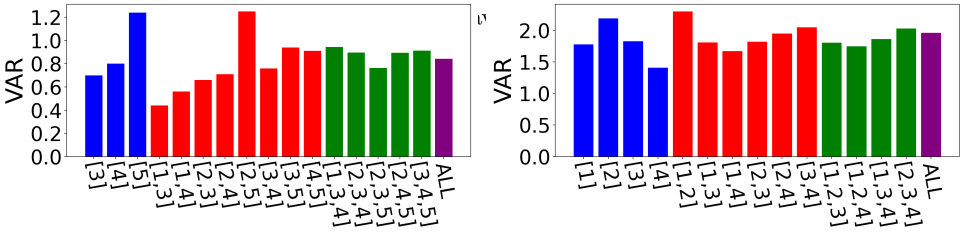
Fig. 3. MJS divergence of different combinations of different confidence level reviews

Divergence Reflected by Jensen-Shannon Divergence. By using euclidean distance, we can only measure the divergence of two sets of different level reviews. Inspired by Jensen-Shannon Divergence for multiple distributions (MJS) [4], we design the MJS metric to measure the divergence between multiple sets of reviews. The MJS of m sets ($m \geq 2$) of different confidence reviews is defined as follows:

$$MJS(l_1, \dots, l_m) = \frac{1}{m} \sum_{i \in \{l_1, \dots, l_m\}} \left(\frac{1}{|R_{l_1, \dots, l_m}|} \sum_{p \in R_{l_1, \dots, l_m}} \overline{s_p^i} \cdot log\left(\frac{\overline{s_p^i}}{\overline{s_p^{[1,m]}}} \right) \right), \quad (2)$$

where $R_{l_1,\dots,l_m}$ is the set of paper IDs, where each paper concurrently has reviews with confidence levels $l_1, \dots, l_m$, $| \cdot |$ returns the size of a set, $\overline{s_p^i}$ is paper p's average review score of l_i-confidence reviews, and $\overline{s_p^{[1,m]}}$ is paper p's average review score of reviews with confidence levels $l_1, \dots, l_m$. The bigger the MJS is, the significant the opinion divergence is. A nice property of MJS metric is that it is symmetric, e.g., $MJS(i, j) = MJS(j, i)$ and $MJS(i, j, k) = MJS(k, j, i)$. We measure the MJS divergence of different combinations of confidence levels and show the results in Fig. 3. Note that, the results of combinations that contain less than 10 reviews are not shown since they are too few to be statistically significant. In 2017-2019 the MJS divergence between conf1 reviews and conf5 reviews is the smallest. In 2020, it shows bigger divergence than 2017-2019 on different combinations but relatively similar divergence results among different combinations. In addition, a combination of three different confidence levels is likely to result in bigger divergence than a combination of two confidence levels.

Divergence Reflected by Average Variance. Opinion divergence also exists within the same confidence-level reviews. The above measurements cannot depict the intra-level opinion divergence. Here, we use average variance to measure the



(a) 2017-2019

(b) 2020

Fig. 4. Average variance of different combinations of confidence levels

intra-level opinion divergence and the inter-level opinion divergence. The average variance of m sets ($m \geq 1$) of different confidence reviews is defined as follows.

$$VAR(l_1, \ldots, l_m) = \frac{1}{m}\left(\sum_{p \in R^n_{l_1,\ldots,l_m}} var(s^1_p, s^2_p, \ldots, s^n_p)\right), \tag{3}$$

where $R^n_{l_1,\ldots,l_m}$ is a set of paper IDs, where each paper concurrently has reviews with confidence levels $l_1, \ldots, l_m$ and the number of reviews with confidence levels $l_1, \ldots, l_m$ is n ($n \leq m$), and $var(s^1_p, s^2_p, \ldots, s^n_p)$ is the variance of paper p's n review scores. Since we will compare the VAR values of different combinations of different confidence level reviews, we have to make sure that the number of samples for variance computation are equal to each other, which can be achieved by introducing the fixed number n. ICLR papers typically have 3 reviews, so we set $n = 3$. The average variance results are shown in Fig. 4. We do not show the results of combinations that contain less than 10 samples. Since there is one more constraint that each combination has to include 3 reviews (refer to the definition of $R^n_{l_1,\ldots,l_m}$), less bars are shown in Fig. 4 than in Fig. 3. In 2017-2019, it is surprised that the maximum variance appears among the most professional reviews (i.e., conf5). While in 2020, the most professional reviews (i.e., conf4) have the minimum variance. It also shows that the variance between the professional reviews and the non-professional reviews is relatively small no matter in 2017-2019 (e.g., conf[1,4]) or in 2020 (e.g., conf [1,4]).

How is the Impact of Non-Expert Reviewers? All these facts demonstrate that the opinion divergence is not greater due to the introduction of more non-expert reviewers. We also observe that the opinion divergence between non-expert reviewers and other reviewers is often relatively small. The reason behind might be that the non-expert reviewers often have a more neutral opinion rather than clear yes or no. They are more cautious to give positive or negative recommendations.

3.2 Which Aspects Play Important Roles in Review Score?

Reviewers often evaluate a paper from various aspects. There are five most important aspects, i.e., novelty, motivation, experimental results, completeness of related workers, and presentation quality. Some conferences provide a peer-review questionnaire that requires reviewer to evaluate a paper from various aspects and give a score with respect to each aspect. Unfortunately, ICLR does not ask reviewers to answer such a questionnaire. Then a question arises accordingly. Which aspects play more important roles in determining the review score? We aim to answer this question by analyzing the sentiment of each aspect.

Corpus Creation. For each review, we first extract the related sentences that describe different aspects by matching a set of predefined keywords. The keywords "novel, novelty, originality, and idea" are used to identify a sentence that describes novelty of the paper, "motivation, motivate, and motivated" are used to identify a sentence related to motivation, "experiments, empirically, empirical, experimental, evaluation, results, data, dataset, and data set" are used to identify a sentence related to experiment results, "related work, survey, review, previous work, literature, cite, and citation" are used to identify a sentence related to the completeness of related work, and "presentation, writing, written, structure, organization, structured, and explained" are used to identify a sentence related to presentation quality. We have collected a corpus containing 95,208 sentences which are divided into five subsets corresponding to the five aspects. Specifically, we have 11,916 sentences related to "novelty", 5,107 sentences related to "motivation", 62,446 sentences related to "experimental results", 8,710 sentences related to "completeness of related work", and 7,029 sentences related to "presentation quality".

Automatic Annotation. In order to train a sentiment analysis model, we need to first annotate enough number of sentences with sentiment label (i.e., positive, negative, and neutral). However, this workload of manual annotation is huge due to the large size of review corpus. Fortunately, we find a possibility of automatic annotation after analyzing the reviews. A large number of reviewers write their positive reviews and negative reviews separately by using the keywords such as "strengths/weaknesses", "pros/cons", "strong points/weak points", "positive aspects/negative aspects", and so on. We segment the review text and identify the positive/negative sentences by looking up these keywords. The boundaries are identified when meeting an opposite sentiment word for the first time. By intersecting the set of positive/negative sentences with the set of aspect-specific sentences, we obtain a relatively large set of sentiment-annotated corpus for each aspect. Particularly, we have 2,893 sentiment-annotated sentences for "novelty", 1,057 for "motivation", 8,956 for "experimental results", 1,402 for "completeness of related work", 1,644 for "presentation quality", and 15,952 in total. We also manually annotate 6,095 sentences including 1,227 corrected automatically annotated sentences since some neutral sentences are incorrectly annotated with positive or negative sentiment. Finally, we have 20,820 labeled sentences[4], i.e., 21.87% of the total number of sentences (95,208) in corpus. Note that, there might be more than one sentences describing one aspect but having different sentiments. In such a case, we label the sentiment by a majority vote.

Sentiment Analysis. Given these five datasets including the labeled data, we perform sentiment analysis for each aspect using a pretrained text model ELEC-

[4] All of the annotated data including manually annotated ones are publicly available at at https://github.com/Seafoodair/Openreview/.

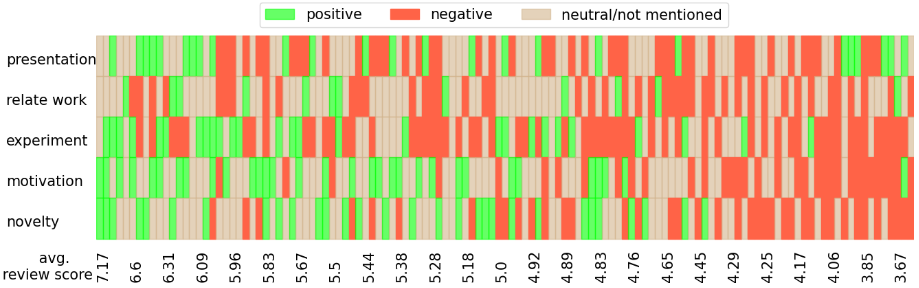


Fig. 5. The sentiment of each aspect vs. the review score. Each column represents a group of reviews with the same combination of aspect sentiments. These groups are sorted in the descending order of the average review score of a group of reviews.

TRA [7] which was recently proposed in ICLR 2020 with state-of-the-art performance. The detailed hyper-parameter settings of ELECTRA are described in our support materials. We split the annotated dataset of each aspect into training/validation/test sets (8:1:1), and use 10-fold cross validation to train five sentiment prediction models for the five aspects. We obtain five accuracy results 93.96%, 88.46%, 94.99%, 85.12%, and 93.38% for novelty, motivation, experimental results, completeness of related workers, and presentation quality, respectively. We then use the whole annotated dataset of each aspect to train the corresponding sentiment analysis model and use this model to predict the sentiment of the other unlabeled sentences of each aspect. Finally, for each review, we can obtain the sentiment score of each aspect. Note that, some individual aspects might not be mentioned in a review, which are labeled with neutral.

Sentiment of Each Aspect vs. Review Score. Given the sentiment analysis results of all aspects of each review and the review score, we perform the correlation analysis. We group the reviews with the same combination of aspect sentiments and compute the average review score of each group. The groups that receive less than 3 reviews are not considered since they have too few samples to be statistically significant. We visualize the result as shown in Fig. 5. We can see that the higher review score often comes with more positive aspects from a macro perspective, which is under expectation. We observe that most of the reviews with score higher than 6 do NOT have negative comments on novelty, motivation, and presentation, but may allow some flaws in related work and experiment. The reviewers that have overall positive to the paper are likely to pose improvement suggestions on related work and experiment to make the paper perfect. The presentation quality and experiment seem to be mentioned more frequently than the other aspects, and the positive sentiment on presentation is distributed more evenly from high-score reviews to low-score reviews. This implies that presentation does not play important role in making the decision. It is also interesting that there is no review in which all aspects are positive or negative. It is unlikely that a paper is perfect in all aspects or has no merit. Reviewers are also likely to be more rigorous in ‘ papers and be more tolerant with poor papers.

Causality Analysis. In order to explore which aspect determines the final review score, we perform causal inference following [10]. Besides the above five aspects, we also include the factor of reviewer confidence. The process of causal analysis includes four steps: modeling, intervention, evaluation, and inference. In the modeling process, we use multivariate linear regression method[3] to perform regression task on the ICLR reviews dataset, where the six evaluated parameters are the sentiment scores of the five review aspects and a reviewer confidence score, and the regression label is the review score. Each parameter is standardized to [-1,1]. To avoid randomness of model training, we launch 1000 times of training and obtain the average MSE (Mean Square Error) 0.24. The intervention process removes each factor x one by one and performs multiple times of model evaluation to obtain multiple average MSE results, each corresponding to an x-absence model. In the absence of overfitting, the MSE value of any x-absence model should be larger than 0.24. The MSE value of the x-absence model implies the causality. A larger MSE value of an x-absence model implies that the factor x is more dominant in determining the final score, and vice versa. In the inference process, we compare the MSE values to infer the causality. The average MSE values of the reviewer confidence, novelty, motivation, experiment, related work, and presentation are 0.84, 0.77, 0.34, 0.86, 0.33, and 0.34, respectively. We observe that the factors of reviewer confidence, novelty, and experiment change the MSE greatly, so they are more dominant in determining the final score.

3.3 Which Research Field has Higher/Lower Acceptance Rate?

AI conferences consider a broad range of subject areas. Authors are often asked to pick the most relevant areas that match their submissions. Area chair could exist who makes decisions for the submissions of a certain research area. Different areas may receive different number of submissions and also may have different acceptance rates. Program chairs sometimes announce the number of submissions and the acceptance rate of each area in the opening event of a conference, which could somehow indicate the popularity of each area. But, the classification by areas is coarse. A more fine-grained classification that provides more specific information is desired. Thanks to the more detailed submission information provided by OpenReview, we utilize the title, abstract, and keywords of each submission to provide a more fine-grained clustering result and gather the statistics of acceptance rate of each cluster of submissions.

We first concatenate the title, abstract, keywords of each ICLR 2020 submission and preprocess them by removing stop words, tokenizing, stemming list, etc. We leverage an AI terminology dictionary [1] during the tokenizing process to make sure that an AI terminology containing multiple words is not split. We then formulate term-document matrix (i.e., AI term-submission matrix) by applying TF-IDF and calculate cosine distance matrix. The size of the term-document TF-IDF matrix for ICLR 2020 is 12436 x 2558, and the size of the cosine distance matrix is 2558 x 2558. We then apply the Ward clustering algorithm [12] on the matrix to obtain submission clusters. Ward clustering is an agglomerative hierarchical clustering method, meaning that at each stage, the pair of clusters with minimum between-cluster distance are merged. We use silhouette coefficient to finalize the number of clusters and plot a dendrogram to visualize the hierarchical clustering result as shown in Fig. 6.

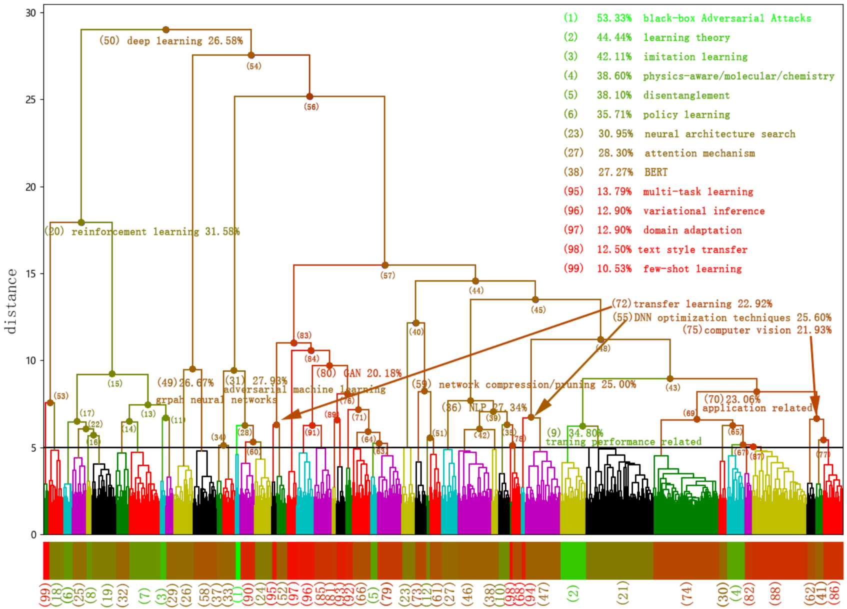


Fig. 6. Visualized hierarchical clustering result of ICLR 2020 submissions. Each leaf node represents a submission. Cosine distance 5 is selected as the threshold to control the granularity of leaf-level clusters. There are 99 clusters in total, including both fine-grained clusters and coarse-grained clusters. Clusters are numbered in the order of their acceptance rate. The color of keywords indicates the acceptance rate of that cluster. Light green means a high acceptance rate, while light red means a low acceptance rate. The keywords of some typical clusters are labeled.

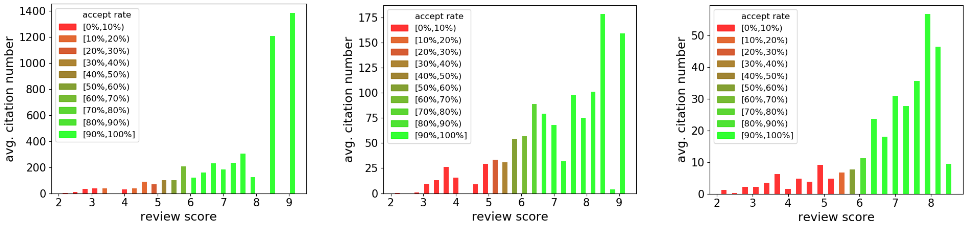
From Fig. 6, we observe three aspects of insights. **(a) Overall Structure of Deep Learning Research**. We observe the correlation between research topics. For example, the submissions in the left part belong to reinforcement learning field (20), which is far apart from all the other research topics (because it is the last merged cluster and its distance to the other clusters is more than 27). Another independent research field is Graph Neural Networks (GNNs) (49), as a promising field, becomes really hot in only 2-3 years, which distinguishes itself from others by focusing on graph structure. Adversarial Machine Learning (31) is also an independent research field that attempts to fool models through malicious input and different from others. The next independent subject is Generative Adversarial Networks (GANs) (80). But GANs is not completely independent since we found that many submissions on NLP (36) and CV (75) are mixed with GANs as well. We also observe that Transfer Learning (72) is close to GANs, since some works have applied transfer learning to GANs. Most of the submissions in the right part are applications related (e.g., vision, audio, NLP, biology, chemistry, and robotics). They are mixed with DNN optimization techniques since many optimizations are proposed to improve DNN on a specific application field. **(b) Popularity Difference between Clusters**. We observe that multiple areas attract large amount of submission. For example, Reinforcement Learning (20), GNNs (49), GANs (80), NLP (36), and Computer Vision (75) have attracted more than 50% of the submissions, which are really hot topics in today's deep learning research. **(c) Acceptance Rate Difference between Clusters**. There exists significant difference on acceptance rate between clusters, say ranging from 53.33% to 10.53%. The cluster of submissions on "Black-Box Adversarial Attacks" has the highest acceptance rate (53.33%), which is a subject belongs to "Adversarial Machine Learning" area. The top-6 highest acceptance rate topics are listed in the figure. The cluster of submissions on "Few-Shot Learning" has the lowest acceptance rate (10.53%), which is a subject belongs to "Reinforcement Learning" area. The top-5 lowest acceptance rate topics are listed in the figure. We also list some typical topics in the figure. For example, the cluster on "Graph Neural Networks (49)" has an acceptance rate of 26.67%. The cluster on "BERT (38)" has an acceptance rate of 27.27%. The cluster on "GANs (80)" has an acceptance rate of 20.18%. The cluster on "Reinforcement Learning (20)" has an acceptance rate of 31.58%.

3.4 Review Score vs. Citation Number

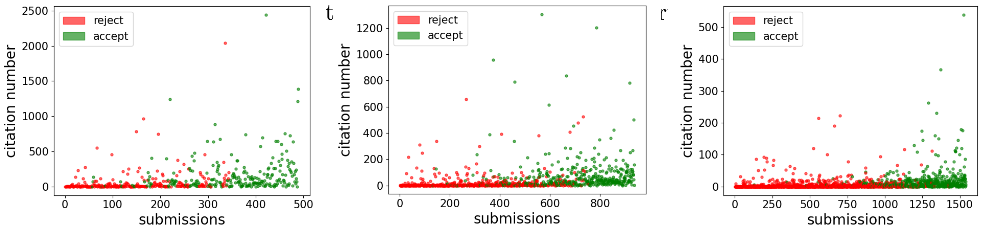
The citation number quantitatively indicates a paper's impact. In this subsection, we show several interesting results on the correlation between review scores and citation numbers.

Is There a Strong Correlation between Review Score and Citation Number? OpenReview releases not only the submission details and reviews of the accepted papers but also that of the rejected submissions. These rejected submissions might be put on arXiv.org or published in other venues and still make an impact. We collect the citation number information of both accepted papers and rejected papers and study the correlation between their review scores and their citation numbers. We plot the histogram of average citation numbers of ICLR 2017-2019 submissions as shown in Fig. 7. The papers are divided into multiple subsets according to their review scores. Each bin of the histogram corresponds to a subset of papers with similar review scores (with an interval of 0.3). Then the average citation number of each subset is calculated. The color of bin indicates the acceptance rate of the corresponding subset of papers. From the figure, we can observe that the papers with higher review score are likely to have higher citation numbers, which is under expectation.

We further investigate the citation numbers of individual papers as shown in Fig. 8. Each point represents a paper. Green color indicates an accepted paper and red color indicates a rejected one. The papers are sorted on the x-axis in the ascending order of their review scores. The distribution of citation numbers is messy. We can see that many rejected papers gain a large number of citations (i.e, red points in the top-left part), which is a bit surprised. Generally speaking, the accepted papers will attract more attentions since they are officially published



(a) 2017 (b) 2018 (c) 2019



(a) 2017 (b) 2018 (c) 2019

Fig. 8. The distribution of citation numbers of individual papers, where the papers on the x-axis are sorted in the ascending order of their review scores

in ICLR. However, the rejected papers may be accepted later at other venues and still attract attentions. In addition, a few papers with high review score are rejected (i.e., red points on the right side). We observe that the reject decision does not impact their citation numbers. Though rejected, the papers with higher review score are still likely to have higher citation numbers.

Do Great Papers Gain More Diverse Review Scores? People always have diverse opinions on the breakthrough works. Reviewer A thinks it novel and is happy to give higher scores, but reviewer B may think it too crazy or unrealistic and reject it. There might be a big debate between reviewers. But it is usually hard to reach consensus. We investigate the relationship between the variance of review scores of a submission and its citation number. We group papers according to their review score variances and calculate the average citation number of each group. Fig. 9 shows the statistical results of the submissions of ICLR 2017-2019. We observe that the papers that have large number of citations are indeed more likely to gain diverse review scores. Note that a paper that has diverse review scores (big review score variance) does not necessarily have high review scores.

3.5 Do Submissions Posted on arXiv have Higher Acceptance Rate?

We found 1,083 submissions that have been posted on arXiv before accept/reject notification[5], which account for about 19.59% of the total submissions. The arXiv versions are not anonymous, which bring unfairness to the double-blind review process. We refer to the submissions that have been posed on arXiv before

[5] We compare paper creation date on arXiv with ICLR official notification date.

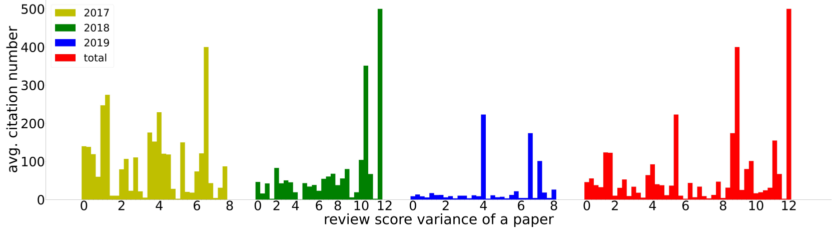
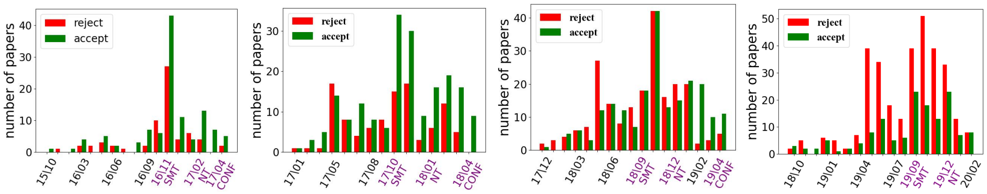


Fig. 9. Review score variance of a paper vs. average citation number

notification as “arXived submissions”. We investigate the acceptance rates of the arXived and non-arXived submissions. The acceptance rates of the arXived submissions in 2017, 2018, 2019, and 2020 are 59.33%, 62.39%, 45.36%, and 30.48%, respectively. The acceptance rates of the non-arXived submissions in 2017, 2018, 2019, and 2020 are 45.88%, 41.23%, 26.37%, and 17.22%, respectively. We observe that the arXived submissions have significantly higher acceptance rate than the non-arXived submissions (49.39% vs. 32.68% on average).



(a) 2017 (b) 2018 (c) 2019 (d) 2020

Fig. 10. The number of accepted/rejected arXived submissions posted on arXiv by month. SMT means the submission deadline. NT means the notification date. CONF means the conference date.

We think the reason should be not only anonymity but also that the arXived ICLR submissions have higher quality. These arXived submissions might attract more feedbacks from colleagues, according to which the authors can improve their manuscripts. The arXived submissions might also be the rejected ones from other conferences and might have been improved according to the rejection reviews. We also observe that some arXived submissions are posted on arXiv one year before the submission deadline. Fig. 10 shows the number of arXived submissions posted on arXiv by month, including both accepted ones and rejected ones. We can see that the papers posted on arXiv are more and more when approaching the submission deadline. There are also a large number of papers posted on arXiv between the submission date and the notification date. From the aspect of acceptance rate, we observe that the earlier the papers are posted on arXiv, the more likely they are accepted. In addition, the papers posted on arXiv after notification date have a higher acceptance rate. The reason might be that the authors cannot wait to share their research results after their papers are accepted.

4 Related Work

There exist many interesting works related to peer-review analysis. We list several related works as follows.

Review Decision Prediction. Kang et al. [13] predict the acceptance of a paper based on textual features and the score of each aspect in a review based on the paper and review contents. They also contribute to the community a publicly available peer review dataset for research purpose. Wang and Wan [20] investigate the task of automatically predicting the overall recommendation/decision and further identifying the sentences with positive and negative sentiment polarities from a peer review text written by a reviewer for a paper submission. DeepSentiPeer [11] takes into account the paper, the corresponding reviews, and review's polarity to predict the overall recommendation score as well as the final decision.

AI Support for Peer-Review System. Anonymous peer review has been criticized for its lack of accountability, its possible bias, and its inconsistency, alongside other flaws. With the recent progress in AI research, many researches put great efforts in improving the peer-review system with the help of AI. Price and Flach [17] survey the various means of computational support to the peer review system. The famous Toronto Paper Matching system [6] can achieve automated paper reviewer assignment. Mrowinski et al. [16] exploit evolutionary computation to improve editorial strategies in peer review. Roos et al. [18] propose a method for calibrating the ratings of potentially biased reviewers via a maximum likelihood estimation (MLE) approach. Stelmakh et al. [19] discuss biases due to demographics in single-blind peer review and study associated hypothesis testing problems.

Other Interesting Works of Peer-Review. Birukou et al. [5] analyzed ten CS conferences and found low correlation between review scores and the impact of papers in terms of future number of citations. Gao et al. [9] predict after-rebuttal (i.e., final) scores from initial reviews and author responses. Their results suggest that a reviewer's final score is largely determined by her initial score and the distance to the other reviewers' initial scores. Li et al. [14] utilize peer review data for the citation count prediction task with a neural prediction model. Hua et al. [13] propose an argument mining based approach to understand the content and structure of peer reviews. The authors leverage the state-of-the-art classifiers to analyze the proposition usage in reviews across ML and NLP venues, showing interesting patterns in proposition types and content. Cormode [8] outlines the numerous ways in which an adversarial reviewer can criticize almost any paper, which inspires us a future work on how to identify the adversarial reviewers based on the open review data.

In this paper, we investigate ICLR 2017-2020's submissions and reviews data on OpenReview and show more different interesting results, e.g., the effect of low confidence reviews, the sentiment analysis of review text on different aspects, the hierarchical relationships of different research fields, etc, which have not been studied before.

5 Conclusion

We perform deep analysis on the dataset including review texts collected from OpenReivew, the paper citation information collected from GoogleScholar, and the non-peer-reviewed papers from Arxiv.org. All of these collected data are publicly available on Github, which will help other researchers identify novel research opportunities in this dataset. More importantly, we investigate the answers to several interesting questions regarding the peer-review process. We aim to provide hints to answer these questions quantitatively based on our analysis results. We believe that our results can potentially help writing a paper, reviewing it, and deciding about its acceptance.

References

1. Artificial-intelligence-terminology (2020), `https://github.com/jiqizhixin/Artificial-Intelligence-Terminology`
2. Some metadata for those curious about their #iclr2020 (2020), `https://twitter.com/cHHillee/status/1191823707100131329`
3. Allison, P.D.: Multiple regression: A primer. Pine Forge Press (1999)
4. Aslam, J.A., Pavlu, V.: Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In: ECIR 2007. pp. 198–209 (2007)
5. Birukou, A., Wakeling, J.R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., Wassef, A.: Alternatives to peer review: novel approaches for research evaluation. Frontiers in computational neuroscience **5**, 56 (2011)
6. Charlin, L., Zemel, R.: The toronto paper matching system: an automated paper-reviewer assignment system (2013)
7. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. In: ICLR 2020 (2020)
8. Cormode, G.: How not to review a paper: The tools and techniques of the adversarial reviewer. ACM SIGMOD Record **37**(4), 100–104 (2009)
9. Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., Miyao, Y.: Does my rebuttal matter? insights from a major NLP conference. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 1274–1290. Association for Computational Linguistics (2019)
10. Gelman, A., Hill, J.: Causal inference using regression on the treatment variable, p. 167–198. Analytical Methods for Social Research, Cambridge University Press (2006)
11. Ghosal, T., Verma, R., Ekbal, A., Bhattacharyya, P.: Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In: ACL 2019. pp. 1120–1130 (2019)
12. Joe, H., Ward, J.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (1963)
13. Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E.H., Schwartz, R.: A dataset of peer reviews (peerread): Collection, insights and NLP applications. In: NAACL 2018 (2018)

14. Li, S., Zhao, W.X., Yin, E.J., Wen, J.: A neural citation count prediction model based on peer review text. In: EMNLP 2019. pp. 4913–4923 (2019)
15. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
16. Mrowinski, M.J., Fronczak, P., Fronczak, A., Ausloos, M., Nedic, O.: Artificial intelligence in peer review: How can evolutionary computation support journal editors? PloS one **12**(9) (2017)
17. Price, S., Flach, P.A.: Computational support for academic peer review: A perspective from artificial intelligence. Commun. ACM **60**(3), 70–79 (Feb 2017)
18. Roos, M., Rothe, J., Scheuermann, B.: How to calibrate the scores of biased reviewers by quadratic programming. In: AAAI 2011 (2011)
19. Stelmakh, I., Shah, N.B., Singh, A.: On testing for biases in peer review. In: ACM EC workshop on Mechanism Design for Social Good. pp. 5287–5297 (2019)
20. Wang, K., Wan, X.: Sentiment analysis of peer review texts for scholarly papers. In: SIGIR 2018. pp. 175–184 (2018)