Reply to IBM Research Zurich's Response

TableMaster Authors

July 4, 2022

In June 30, 2022, the authors of the TableFormer [1] from IBM Research Zurich have posted a response letter on https://research.ibm.com/blog/tableformer-response, entitled "Response to accusations regarding TableFormer paper". Disappointingly, the responses are full of misleading or even self conflicting contents to confuse the audiences. Thus, we, the authors of the TableMaster [2], have to prepare a set of point-to-point technical replies to clarify the truth.

First of all, we would like to refer the audiences to the standard definition about "plagiarism" from the Oxford University[1].

Definition 1 *"Plagiarism" is defined as the copying or paraphrasing of other people's work or ideas into your own work without full acknowledgment. All published and unpublished material, whether in manuscript, printed or electronic form, is covered by this definition. "Collusion" is another form of plagiarism involving the unauthorised collaboration of students or other individuals in a piece of work.*

1 Summary Reply to IBM Research Zurich's Response

1.1 *The Summary Response from IBM Research Zurich*

- *"We, IBM researchers and the authors of the TableFormer [1] work, would like to respond to accusations of plagiarism by the authors of TableMaster [2] (later referred as "OP") in regard to their ideas and code."*
- *"The accusations arose on June 27, 2022 following the publication of our paper at the Computer Vision and Pattern Recognition Conference (CVPR). The authors of TableMaster did not contact us prior to their public accusations, which are ungrounded and easily refuted by a simple comparison of the two papers in question."*
- *"First, though, we would like to point out that never, in this or any other instance, have IBM researchers plagiarized anyone's work. We adhere to the highest ethical standards of research and publishing our work, be it as a pre-print, at a conference or any other venue, or in a journal."*

[1] https://www.ox.ac.uk/admissions/graduate/applying-to-oxford/university-policies/plagiarism

- *"Our work introduces a different neural network architecture, built on top of the work published in 2019 by our IBM colleagues (EDD [3]). Also, TableFormer uses a unique data processing pipeline, applied directly to programmatic PDF documents. This approach follows our 2018 KDD [4] paper ideas for PDF parsing and is fundamentally different from TableMaster, which depends on Optical Character Recognition (OCR) from images."*

1.2 Summary Replies from TableMaster (Ours)

We, the authors of the TableMaster [2], would like to point out the following facts.

1. IBM Research had organized a competition[2] during July 20, 2020 to May 1st, 2021, and many companies and universities have submitted their solutions, including our TableMaster [2]. Our submitted solution TableMaster [2] has been ranked the 2nd place. Then we posted our preprint paper on arXiv (arXiv: 2105.01848, `https://arxiv.org/abs/2105.01848`), the implementation code for the model (July 29, 2021), demo, slides, docker files, and etc. **Coincidentally, an extremely similar solution—TableFormer [1]—by the IBM Zurich researchers then was submitted to CVPR 2022 (six months later) and published (13 months later).**

2. IBM Zurich researchers said that they were not even aware of our work. TableFormer [1] cited the official technical report for the competition [5]. The technical report included the results of 9 solutions, but TableFormer [1] quoted merely one result. Usually, it is obliged for the authors to cite many relevant works in order to give a comprehensive comparison and to show the superiority of their proposal. It is weird to ignore our submitted solution—which yielded the second-best result in the competition—but the IBM Zurich researchers then submitted an extremely similar solution to CVPR, in Nov. 2021. **Is such a weird and suspicious scheme to produce paper representing the "the highest ethical standards" of research and publishing your work"? If "yes", maybe either "highest" or "ethical standard" needs to be redefined in some way.**

3. It is claimed that TableFormer [1] was built on top of the work published in 2019 by the IBM colleagues (EDD [3]). However, EDD is unlikely relevant to the current solution in TableFormer [1]. EDD [3] is a simple end-to-end solution for table recognition. TableMaster and TableFormer both use the strategy of "divide and conquer". **To be specific, TableMaster introduced a multi-stage processing method including table structure regression, complex post-processing, and etc.; whereas TableFormer used an extremely similar method,** which will be discussed in details in the next part.

4. It is claimed that TableFormer [1] used an unique data process processing pipeline to extract PDF cells for a programmatic PDF documents. In this task, the motivation of extracting "PDF cells" is similar to using text line detection and text line recognition in our TableMaster [2]—both of them is to obtain the content and the position of the content, and then to serve the subsequent matching pipeline (post-processing). It is quite weird to directly extracting such information! Suppose that the table structure and content can be extracted directly from a digital PDF by a PDF parser, as the authors of the TableFormer [1] claimed

[2] `https://icdar2021.org/program-2/competitions/competition-on-scientific-literature-parsing/`

in the response, then what is the meaning of table recognition? Please notice that in the dataset PubTabNet, which is provided by the organizer (IBM) to the public participants, the table to be processed is in the format of ".PNG"! Can you really extract the "PDF cells" directly by a PDF parser? Or did you use the original ".PDF" format to obtain the ground truth information of the content? **If yes, what is the fairness of academic research? It is so weird.**

Now, we would like to give an introduction to the mentioned three methods: EDD [3], TableMaster [2] and TableFormer [1].

- EDD (first released in Nov. 2019) used an end-to-end method to directly estimate the table (including the structure and content in each cell). In their network, they used two classification branches: one for the structure and one for the content. Note that no cell box regression was needed or used in EDD, thus no regression loss is used. In fact, EDD did not need a complex post-processing. We know that the dataset PubTabNet contains many complex tables (the length of the sequences of the table, including the structure and content, is extremely long, maybe 5,000 or more). The solution in EDD is too simple to be employed for the complex real applications.

- TableMaster (our method, firstly released in May 2021) used a four-stage method. We converted the table recognition into four sub-tasks, including table structure recognition (cell prediction and box regression), text line detection, text line recognition, and matching pipeline between the table structure and text line content. One key innovation in our method was to recognize the table structure by jointly predicting the structure cell and regressing its bounding boxes. Meanwhile, to serve the whole process, we introduced some complex but effective post-processing rules.

- TableFormer (from the IBM researchers, first released in March 2022) used a three-stage method, including table structure recognition (cell prediction and box regression), PDF cells extraction (Our solution: employing a text line detection and a text line recognition; IBM Zurich researcher's solution: using PDF parser to directly obtain the ground truth of the text line content), and matching pipeline between the table structure and text line content. For table structure recognition, TableFormer also used a joint prediction of the structure cell and regression of the bounding boxes of the cells—this is extremely similar to ours. **Given the prior work in our TableMaster, could you tell what is essentially novel in your TableFormer?**

All the previous methods other than TableFormer, including EDD, Davar-Lab-OCR [5], our TableMaster [5], TAL-solution [5], PaodingAI-solution [5], are designed to extract the content from the digital images (".PNG" format). Then, you said you used the PDF parser to directly obtain the content. Did you mean that you used the ground truth, but other researcher used OCR technique to recognize the content?

Finally, we would like to point out that the ethical standards of a company should be accessed by the audiences, and by the academic community rather than its own claims.

2 Our Replies to the Response for Denying Idea Copy

2.1 *Response from IBM Research Zurich—"TableFormer: Did we copy the idea? The answer is no."*

- *The dual decoder approach was introduced by our colleagues at IBM in 2019 before the OP's work (in 2021).*
- *The EDD4 public code*[3] *contains the idea of bounding box regression, which predates the code-base and paper of the OP. In our quantitative analysis section, we refer to it as "EDD+BBox.".*
- *The TableFormer network architecture is different from TableMASTER-mmocr. TableMASTER-mmocr uses a dual transformer decoder, text-line detection (based on PSENET). But Table-Former uses a single transformer decoder, with the output of the transformer decoder being first used by an attention network, and then with a DETR [6] head to predict the bounding box.*

2.2 Our Reply

To clarify the truth, we would like to point out the following points.

- In our previous accusations, we said "Our key idea in Tablemaster is to formulate the Table structure recognition into a joint box regression and token classification, the network structure in the plagiarized TableFormer is exactly the same as our Tablemaster." We did not mention that the dual decoder approach was firstly introduced by ours, and we believe that many previous methods used the way of two classification branches. **We said that one of our key innovations was to formulate the table structure recognition into a joint box regression and token classification problem. And it is weird that the TableFormer used exactly the same method.** As we mentioned above, your colleagues' method, EDD (first released Nov. 2019), used simply an one-stage method to directly estimate the table (including the structure and content) and thus, it will encounter big issues when the table is complex.
- Let us search in the EDD paper, we cannot find any term mentioning "box regression", "regression", "ℓ_1 loss", "ℓ_2 loss". The method of EDD has nothing to do with our method (a joint box regression and token classification). In fact, even in the EDD code you provided in its link, only an invalid function was defined in the code, and that the code was not used in any loss computation, training and inference. The code is disabled in default.
- You said your method was motivated by DETR [6] (Please cite the correct paper. It is DETR, please not citing to deformable DETR!). The two key innovations in DETR are: a) introducing a set of learnable queries and b) introducing a set-based global loss computed by Hungarian bipartite matching. **We wonder: which one did you actually use? Be honest, TableMaster and TableFormer were based on Transformer, but not on DETR.** Could you find the difference between Transformer and DETR?

[3] https://github.com/ibm-aur-nlp/EDD/blob/5fbbe5b9473953528296befa1c3e1ae868a36b74/models.py#L1356

3 Our Replies to the Response on Denying Model Copy

3.1 *Response from IBM Research Zurich—"TableFormer: Did we use any of the OP's models? The answer is no."*

- *We do not use OCR—instead, we use the content from the original PDFs.*
- *We do not use OP's "text line detection" or "text line recognition." In fact, we do not need to do this process at all, because we do not use any OCR*
- *We only use the original PDFs[4] developed by our colleagues to create the PubTabNet dataset.*
- *We apply our own method, published in 2018, to extract content (raw files) from PDFs.*

3.2 Our Reply

We would like to make the following clarification.

- The standard pipeline is to detect the text line and recognize the text line content. Then, you said you can extract the ground truth of the text line by a PDF parser. Then, what is meaning of table recognition? The data you provide to the public is the ".PNG" image, and then you said you use a PDF parser to directly extract the ground truth content from the digital PDF file (you synthesis). What is the fairness of academy research? Interesting!

4 Our Replies to the Response to Denying on Visualizations

4.1 *Response from IBM Research Zurich—"TableFormer: Did we use any of the OP's visualizations? The answer is no."*

- *Using bounding boxes to visualize detections is a standard technique in computer vision.*
- *Many papers, published before OP's work, use bounding boxes to visualize detections in tables. One example is the work [7] by our IBM colleagues in 2020.*
- *Our visualization is produced using our Javascript/HTML code, which has a unique appearance and simplifies the comparison of predictions at different stages.*

[4] `https://ibm.ent.box.com/s/a26ofti6bqd77zsah6xoe3d8oryztypv`

4.2 Our Reply

- The visualizations of TableFormer were similar with the visualizations of TableMaster. Considering so many same or similar points between TableFormer and our TableMaster, It is hard for us to believe that the TableFormer did not refer to our TableMaster.

5 Our Replies to the Response on Denying the Pre-processing

5.1 *Response from IBM Research Zurich—"TableFormer: Did we copy the OP's pre-processing? The answer is no."*

- *Our data preparation stage includes steps that are not present in OP's work. For instance, we have introduced a procedure that generates the missing bounding boxes explained in our supplementary material.*
- *In the implementation details of our paper, we explained why we used 512 tokens.*
- *The HTML classification tokens are not defined by OP's work, but they were first described by EDD in 2019.*
- *Even OP's screenshots show that our work is different than theirs because we use "uncollapsed" tokens ("⟨td⟩", "⟨/td⟩") — contrary to their work that uses "collapsed" tokens ("⟨td, /td⟩").*

5.2 Our Reply

- According to your responses, it is obvious that you are so familiar with your colleagues' works. Now, your colleagues' competition technical report have reported top-9 best performing methods, but you only cited one result (you quoted exact result from the report)—is it just because our result were the second best method?
- You said you use "uncollapsed" tokens (individual, not merge tokens) instead of our merged tokens, we have conducted a statistic to the PubTabNet dataset. With the unmerged token sequence, the sequence lengths of around 3.9% tables are longer than 512. It means the performance upper bound should be below than 96.1%, and even in the TEDS index, the performance upper bound should be below than 97%, how can you get your result 96.75%. Please carefully read your reported performance and the results you compared, there are many errors and misleading information.
- How did you compare your results with other methods (see below tables copied from different papers)? Is it a fair comparison? Your results are obtained by using the ground truth of the table content; whereas other methods need to use OCR technique to recognize content.

It is obvious that the authors fully understood the competitions and they cited the technical report and even mentioned the reported performance in the report. However, the results are intentionally misleading or confusing. Let us see the results (top-9 results in the competition) from

the competition technical report. For convenience, we copied them from the technical report [5] and displayed them in Figure 1 (c). For a clear comparison, we also copied the results from TableFormer [1] and displayed them in Figure 1 (a) and (b).

Then, let us see their corresponding results in their CVPR manuscript, that is claimed in the "highest ethical standards" of research. They are shown in Figure 1 (a) and (b). The results listed in Table 4 at panel (a) and Table 2 at panel (b) are problematic, intentionally misleading and confusing reviewers and audiences. In panel (a), the listed baseline results are just a few "straw-man". To have a fair and correct comparison:

- The results in blue box of Table 2 at panel (b) should be put into the Table 4 at panel (a); and
- The results from the competition results [5] in Table 4 at panel (c) should be included.

Moreover, the results of TableFormer [1] in Table 2 at panel (b) are also problematic due to unfair settings to evaluate the performance. To be specific, all results in Table 4 at panel (c) are evaluated with respect to the whole table content—including both table structure and content; whereas the results of TableFormer [1] and other results in Table 2 at panel (b), except for Davar-Lab, are evaluated with respect to the table structure only. It is unfair to mix them into a single table, which is misleading the performance comparison. Could you please tell the reason why and how such a mass tables are produced? Does it demonstrate your claimed "highest ethical standards" of research and publishing work? Interesting!

Their reported performance in the bottom row (boxed in red) of Table 4 in Figure 1 panel (a) cannot even beat any method in Table 4 at panel (c) which are from competition report [5]. **Note that the worst results are still 94.84%, better than TableFormer's result 93.6%. Then, it is very interesting to claim that "significantly outperforms SOTA" by more than 5%. What is your definition of "SOTA"?**

6 Ours Replies to the Response on Denying Copy Post-processing

6.1 Response from IBM Research Zurich—"TableFormer: Did we copy the OP's post-processing? The answer is no."

- *Our TableFormer extracts the text directly from the PDF document and it does not use any OCR. Therefore, the output of our model is different and uses different post-processing treatment.*
- *Out post-processing pipeline is more sophisticated than OP's work. This has been explained in detail in our supplementary material.*
- *Caching for auto-regressive methods during the inference is a known practice. It has been implemented by Open-Source Neural Machine Translation (OpenNMT) and is described in this blog post*[5].

[5] https://scale.com/blog/pytorch-improvements

Model	TEDS		
	Simple	Complex	All
Tabula	78.0	57.8	67.9
Traprange	60.8	49.9	55.4
Camelot	80.0	66.0	73.0
Acrobat Pro	68.9	61.8	65.3
EDD	91.2	85.4	88.3
TableFormer	95.4	90.1	**93.6**

Table 4: Results of structure with content retrieved using cell detection on PubTabNet. In all cases the input is PDF documents with cropped tables.

(a) Copied from TableFormer [1]

Model	Dataset	TEDS		
		Simple	Complex	All
EDD [38]	PTN	91.1	88.7	89.9
GTE [37]	PTN	-	-	93.01
Davar-Lab [13]	PTN	97.88	94.78	96.36
TableFormer	PTN	98.5	95.0	**96.75**
EDD [38]	FTN	88.4	92.08	90.6
GTE [37]	FTN	-	-	87.14
GTE [37] (FT)	FTN	-	-	91.02
TableFormer	FTN	97.5	96.0	**96.8**
EDD [38]	TB	86.0	-	86.0
TableFormer	TB	89.6	-	**89.6**
TableFormer	STN	96.9	95.7	96.7

Table 2: Structure results on PubTabNet (PTN), FinTabNet (FTN), TableBank (TB) and SynthTabNet (STN).
FT: Model was trained on PubTabNet then finetuned.

(b) Copied from TableFormer [1]

Team Name	TEDS Simple	TEDS Complex	TEDS all
Davar-Lab-OCR	97.88	94.78	**96.36**
VCGroup	**97.90**	94.68	96.32
XM	97.60	**94.89**	96.27
YG	97.38	94.79	96.11
DBJ	97.39	93.87	95.66
TAL	97.30	93.93	95.65
PaodingAI	97.35	93.79	95.61
anyone	96.95	93.43	95.23
LTIAYN	97.18	92.40	94.84

Our results

Table 4. Task B top TEDS results. The overall result (TEDS all) is decompose into simple and complex tables [16]

(c) Copied from competition report [5] (which is released by May 1st, 2021)

Figure 1: Comparison on Reported Results in Different Papers. The reported results in the bottom row (boxed in red) of Table 4 at panel (a) from TableFormer [1] should be compared to the listed top-9 results in Table 4 at panel (c) which are from competition report [5]. It is clear that the reported results of TableFormer [1] is inferior than the results from any result in the competition report [5]. The results in Table 2 at panel (b) are essentially evaluated with different and unfair settings. How could you claim the reported results being "significantly outperforms SOTA"? Interesting!

6.2 Our Reply

- Post-processing itself is an independent stage. It doesn't depend on whether using the results of the text detection plus recognition method or the results of directly extracting PDF cells.
- The post-processing pipeline was firstly proposed in our TableMaster to serve our proposed multi-stage approach. We have introduced three complex rules, then you broke down our three rules into nine trivial points—each point can be classified into one of these three rules. **Then you said yours were more sophisticated than ours! Is there any essential difference?**
- For the memory cache inference, we have pointed out that we were inspired by XLNet and we were very early in applying it in the OCR community. Given so many mentioned facts above and that the table recognition is a very specific and very small sub-field in OCR, it is really hard for us to believe what you claimed in the misleading response.

7 Ours Replies to the Response on Denying All

7.1 Response from IBM Research Zurich—"TableFormer: Did we mislead anyone? The answer is no."

- *We were not aware of the OP's work. Even during the paper's review process, the existence of OP's work was not mentioned.*
- *The OP did not contact us prior to a mass email sent to our work colleagues, and the Reddit post with the accusations. It would have been better if the OP had contacted us before making public accusations, then we could have gladly proven our points, cited OP's work and compared the approaches.*
- *We are open for discussion with the OP to further clarify all of the above and prove that our work has not been copied or even inspired by the OP's work.*
- *We demand a retraction of the plagiarism accusation, and an apology email to our colleagues retracting the accusation.*
- If the OP is still not convinced, we don't mind them reaching out to CVPR. We have overwhelming evidence in terms of code (git history) and documentation to prove that the accusations are completely and totally unfounded and false.

7.2 Our Reply

- We think that the audiences, the experts in the relevant field, can make a judgement by themselves: a) if the TableFormer authors are aware or not aware of our previous work, and b) if TableFormer was built on our TableMaster or their colleagues' work EDD.

- This is an academic integrity issue, not a citation issue. We made the public accusations not because we wanted to be cited, but because we were copied intentionally.
- We have submitted the plagiarism to the CVPR community and the IEEE. We hope they will set up an ethics committee and invite some experts in the field of OCR or table recognition to investigate and evaluate this event. At the same time, we also hope that the IEEE can make a clear definition of plagiarism or academic misconduct for future CV or AI conferences.

References

[1] A. Nassar, N. Livathinos, M. Lysak, P. Staar, Tableformer: Table structure understanding with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4614–4623.

[2] J. Ye, X. Qi, Y. He, Y. Chen, D. Gu, P. Gao, R. Xiao, Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: Table recognition to html, arXiv preprint arXiv:2105.01848 (2021).

[3] X. Zhong, E. ShafieiBavani, A. J. Yepes, Image-based table recognition: data, model, and evaluation, arXiv preprint arXiv:1911.10683 (2019).

[4] P. W. Staar, M. Dolfi, C. Auer, C. Bekas, Corpus conversion service: A machine learning platform to ingest documents at scale, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 774–782.

[5] A. Jimeno Yepes, P. Zhong, D. Burdick, Icdar 2021 competition on scientific literature parsing, in: International Conference on Document Analysis and Recognition, Springer, 2021, pp. 605–617.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213–229.

[7] X. Zheng, D. Burdick, L. Popa, X. Zhong, N. X. R. Wang, Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 697–706.