MARGo

Multi-Agent Review Generation for Scientific Papers with gpt-4o

Alexandra Elbakyan

abstract

Peer-review is a central part of decision making process in scientific journals. Current peer-review system has a number of disadvantages, such as being susceptible to bias; furthermore, as the number of papers submitted to academic journals increases, it presents additional challenge for journals to process all submissions they receive. Advances in AI and NLP can potentially enable automatic processing and evaluation of research manuscripts, significantly reducing the amount of human work. One of the recent developments in that domain is MARG, a multi-agent system for automatic review generation based on GPT-4. The latest version of MARG has a number limitations, such as lack of support of recent OpenAI models and interfaces, or outdated code organization practices that largely prevent reproduction of results and further investigation and development. This project addresses these shortcomings by presenting a new framework based on MARG that makes research and experiments in multi-agent scientific review generation significantly more accessible, as well as engaging and therefore suitable for use in educational setting.

Project overview

Research Summary: In this project AI will be applied to the task of automatic generation of peer-reviews for scientific articles. The problem is itself new, as for many decades it remained out of reach for conventional machine learning methods. The past few years have witnessed a number of breakthroughs in AI, including development of the field of generative AI and Large Language Models. These new developments have greatly expanded the scope of problems that can be solved with AI methods. For processing large and complex texts, such as scientific articles, particularly important was development of novel transformer neural network architectures, that enabled large-scale language modeling. Not only these neural networks allow for a more nuanced undertanding of human language, but also retain large amount of knowledge when trained on vast arrays of textual data.

The possibillity of using LLM for automatic peer-review has been investigated in a number of recent studies. These studies primarily focus on GPT model from OpenAI, currently the most advanced model available for public usage. In standard scenario, a single language model is prompted to generate summary of a research study. A few publications explore a more advanced approach, involving multiple AI agents interacting between each other and discussing the paper. Given that peer-review is inherently a collective process, that approach can potentially yield better results. The most advanced system for multi-agent review generation currently available is MARG, developed by Allen AI Institute.

Modifications & Enhancements: MARG system has a number of disadvantages that prevent it from being widely used in research and education and improved further upon, such as script-based code organization and lack of interactivity. The aim of the current project is to provide a new framework for multi-agent review generation based on MARG, that will be highly interactive, customizable and extensible, making it easy to design and run various experiments.

Implementation and Results: The resulting framework, MARGo, implements several classes that can be used for building multi-agent systems for simulation of scientific peer-review process. The system provides additional features, such as monitoring discussions between agents, that make the process engaging, which is important in e.g. educational settings.

introduction

Peer-review of scientific papers is a central part of decision making process in academic journals. A research paper submitted to the journal is typically assigned 2-3 reviewers with an expertise in corresponding fields, who are requested to provide detailed responses regarding originality and impact of research study, its potential flaws and areas for improvement. Based on results of peer review, editor makes decision either to publish paper in a journal, reject it, or request authors to do additional work and improve / revise the paper.

Results of peer review process can be highly subjective and susceptible to bias. To eliminate bias, some journals have introduced double-blind mode, when author identity is not revealed to reviewers. Studies have shown that under double-blind peer-review, acceptance rate is increased for female authors and decreases for famous authors and authors from high-prestige institutions [1-8]. However, double blind peer-review cannot completelty eliminate human factor and guarantee anonymization of author identity in e.g. small research areas. Furthermore, an increasing number of research works today are being made available open access before publication, in a preprint form.

In addition to this, the number of subsmissions to academic journals has been continuously growing due to acceleration of scientific progress, making in harder for the journal to recruit enough professional reviewers to handle that volume [9].

Automatization of peer review process can be a potential remedy to these problems. Although machine learning and natural language processing methods today are already used by academic publishers for automatic pre-screening of submissions [10], the area is still in its infancy and major part of the review process remains non-automated. Recently, large language models have revolutionized the field of AI. Neural networks based on transformer architecture provide a much more nuanced understanding of human language, being able to memorize a large amount of knowledge from unstructured text while training and use it to asnwer questions and generate long human-quality responses on various topics. These qualities potentially make language models suitable for such complex tasks as reviewing research manuscripts.

related work

The ability of large language models, primarily ChatGPT, to generate reviews for academic papers has been explored in a number of recent studies. However, most of them report that reviews written by ChatGPT do not correlate with those received from human revewers [11], the LLM is unable to predict the outcome of peer-review process, i.e. acceptance or rejection [12] and fails to recognize major inconsistencies that were deliberately introduced to articles for testing [13]. Regardless of this, many peer-reviewers today apparently use ChatGPT [14-16]. For example, a recent study [14] estimated that around 16% of reviews at AI conference in 2024 have been using AI tools in writing, which is noticeable by increasing number of adjectives that are often used by ChatGPT [15]. However, it remains an open question, was ChatGPT used to generate the whole review, or check the grammar only. Interestingly, papers reviewed by AI were more likey to be accepted.

A study conducted by Verharen e.al. has shown ChatGPT to successfully identify gender disparities in tex [17]t. Studies [18,19] employ multi-agent framework to simluate peer review process. This approach will be analyzed in more detail in the next section.

MARG

MARG is a system for multi-agent review generation developed by Allen AI Institute in 2023 [19]. The initial problem MARG aimed to solve was limitations on the size of the context window in early version of GPT-4, that was limited to 8192 tokens - not enough to process a whole research paper. To approach this problem MARG used multiple AI agents, each processing different chunk of the paper. Reviews generated by individual agents are then summarized by expert and main agents. The system was evaluated and shown to increase the amount of usable or actionable feedback in reviews, versus generic comments typically produced by ChatGPT barebone.

Since MARG was introduced, large language models has evolved, and latest version of GPT now supports up to 128K tokens in the context window [20]. Recently, Titan architecture for neural networks was presented by Google, thar enabled even larger context size up to 2M [21]. Still,

multi-agent review generation can have advantages, as it enables different AI agents to focus on specific sections and aspects of the paper, producing specific commentaries rather than general overview of the study.

Current version of MARG availabe for public download on GitHub [22] has a number of limitations that significantly affect its usability and further development. Even though these limitations are intertwined and cannot be rigorously separated from each other, some key issues that can be named are following:

1. Lack of support for the most recent models from OpenAI such as gpt-4o. The name of the old model is hard-coded in multiple places within the code, but changing the name is not enough, since output format is also somewhat different, so the code will not work out of the box but crash instead.
2. The system is designed for end user rather than researcher, and must be accessesed through a web interface. The interface does not provide any configuration options to experiment with, such as name of the model, temperature or task/instruction prompts. User is required to upload PDF file and then wait for the system to produce result by running different review generation algorithms described in the MARG paper. The process takes considerable amount of time without any possibility to control the flow, interact with agents or monitor progress — altough some log output is provided in the console, it is not formatted and does not allow to clearly understand the activity. When the task is done, only final review texts will be provided as output. For research and educational purposes, having access to full discussion between different agents (editor, experts and reviewers) would be beneficial, as it allows to better understand the system and potentially improve it.
3. The code follows bad programming practices and is not convenient for researcher as well. Commentaries are lacking, as well as explanations of the architecture and purpose of available classes and methods, which makes it practically impossible to re-use the code for different tasks or experiments.
4. MARG is built upon traditional chat completion API, while OpenAI has recently introduced Assistants API specifically designed for implementing various kinds of agents [23].

There definitely exist some journal editors who will be interested using automatic generation of reviews for submitted papers as an experimental approach; researchers themselves could run MARG over their works in order to understand weak spots and improve them before sending the manuscript out for peer-review. However, given the fact that technology is still in its early stages, the main interest should be expected from AI researchers and graduate students. That emphasizes the need for a much more flexible, interactive and extensible framework, that will implement core principles of MARG while at the same time enabling anyone to design and run experiments and test and implement various ideas in multi-agent processing of scientific texts.

methodology

To generate review texts, this project implements a general algorithm described in MARG paper. First, a research paper is converted to text using GROBID service. The text is then split to N chunks, and N reviewer agents are initialized to analyze every chunk. In addition to that, leader and expert agents are created. Leader is orchestrating the review process, handles communication with the user and is responsible for writing the final version of the review based on feedback received from reviewer agents, as well as from expert. The latter does not have a chunk of the paper assigned, and recieves neccessary information from the group leader. The task of an expert is to scrutinize the article and come up with possible questions.

To improve upon original MARG project, MARGo is implementing a novel *interactive programming* methodology, that organizes code, results, graphics and text within a cells of a single document. In the past few years, interactive programming within Ipython / Jupyter notebook environment has emerged as a standard in data science [24]. The approach has numerous advantages over traditional script-based programming implemented in MARG, enabling greater modularity and reproducibility and simplifying collaboration and code re-use. However, not every researcher has accepted the novel approach: apparently, engineers at the Allen Institute for AI did not endorse using notebooks [25] and that is probably the main reason why MARG was built with classical scripting approach.

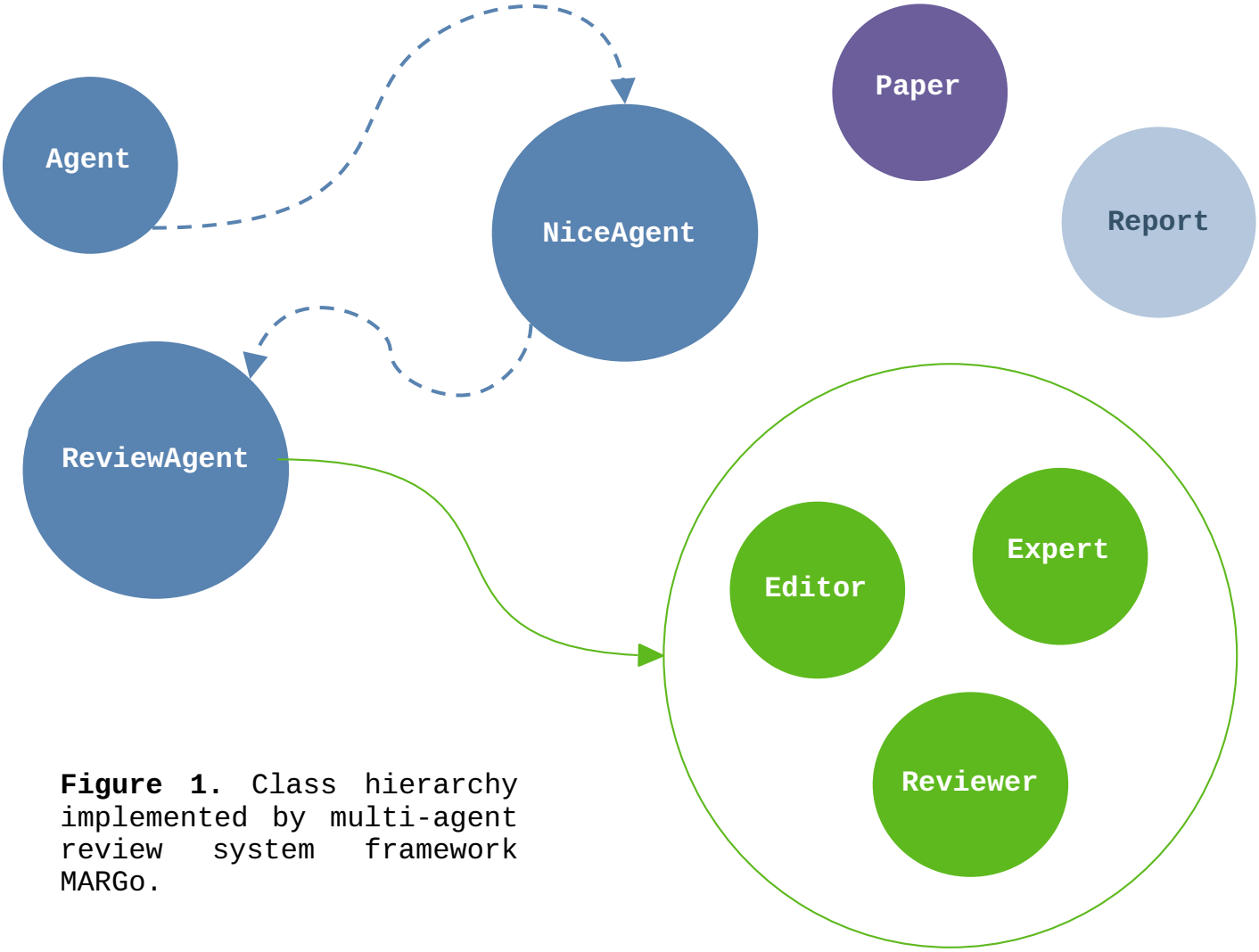

Figure 1. Class hierarchy implemented by multi-agent review system framework MARGo.

The code has modular structure and implements several classes that can be re-used for building a custom multi-agent system (Figure 1). The basic class Agent handles communication with OpenAI servers using new Assistants API. By default, gpt-4o model is used. The class supports agents to have name and avatars; if these was not provided on initialization, name will be randonmly generated and avatar selected from the photo bank. NiceAgent improves upon basic class, extending it with streaming and formatting support.

The review process is performed bu Editor, Expert and Reviewer classes. Only the Editor class must be instantiated. When paper is submitted to Editor, it will recruit an expert and N reviewers according to the number of paper chunks, and start the review process.

Paper class provides functionality to load and pre-process the paper, extracting text and cutting it into chunks. Report class provides methods to export review and discussion to HTML and JSON files.

results

Example output of the system is shown in the Appendix. The task was to review classic paper about Transformers «Attention is all you need» [26]. To perform the task, editor agent had to recruit five reviewers and one expert.

The following is a summary of improvements / modifications of the original MARG generator that were implemented:

1. Using new Assistant API in place of completion API.
2. Support for the latest gpt-4o model.
3. Interactive programming paradigm implemented, the system is running within a Jupyter notebook and does not require web server.
4. The code is commented in detail, and logically organized into a few clases that can be re-used to design and text custom MA systems.
5. Agents are given a name, and the name is used in communication protocol, where the «From:» section was added. Each of the reviewers and expert are given the name of the group coordinator (in the original MARG systems, names are not used)
6. Editor agent is instructed to print out a stop-word «READY» when review is finished, to avoid infinite loops that sometimes happen in MARG. In addition to that, when submitting the paper, maximum number of comments can be restricted (default 128)
7. MARGo implements formatted and live display of answers and whole discussion streams, updated as soon as new answer or new token is available. Each agents is also given an avatar (profile picture) that is displayed next to the answer. Silent mode for running batch jobs is also supported.
8. The result of the peer-review (discussion stream) can be saved to JSON and HTML output file.

references

1. R. M. Blank, “The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review,” The American Economic Review, vol. 81, no. 5, pp. 1041–1067, 1991.
2. A. E. Budden, T. Tregenza, L. W. Aarssen, J. Koricheva, R. Leimu, and C. J. Lortie, “Double-blind review favours increased representation of female authors,” Trends in Ecology & Evolution, vol. 23, no. 1, pp. 4–6, Jan. 2008, doi: 10.1016/j.tree.2007.07.008.
3. T. J. Webb, B. O’Hara, and R. P. Freckleton, “Does double-blind review benefit female authors?,” Trends in Ecology & Evolution, vol. 23, no. 7, pp. 351–353, Jul. 2008, doi: 10.1016/j.tree.2008.03.003.
4. E. S. Darling, “Use of double-blind peer review to increase author diversity,” Conservation Biology, vol. 29, no. 1, pp. 297–299, 2015.
5. K. Okike, K. T. Hug, M. S. Kocher, and S. S. Leopold, “Single-blind vs Double-blind Peer Review in the Setting of Author Prestige,” JAMA, vol. 316, no. 12, pp. 1315–1316, Sep. 2016, doi: 10.1001/jama.2016.11014.
6. A. Tomkins, M. Zhang, and W. D. Heavlin, “Reviewer bias in single- versus double-blind peer review,” Proceedings of the National Academy of Sciences, vol. 114, no. 48, pp. 12708–12713, Nov. 2017, doi: 10.1073/pnas.1707323114.
7. A. R. Kern-Goldberger, R. James, V. Berghella, and E. S. Miller, “The impact of double-blind peer review on gender bias in scientific publishing: a systematic review,” American Journal of Obstetrics and Gynecology, vol. 227, no. 1, pp. 43-50.e4, Jul. 2022, doi: 10.1016/j.ajog.2022.01.030.
8. M. A. Ucci, F. D’Antonio, and V. Berghella, “Double- vs single-blind peer review effect on acceptance rates: a systematic review and meta-analysis of randomized trials,” American Journal of Obstetrics & Gynecology MFM, vol. 4, no. 4, p. 100645, Jul. 2022, doi: 10.1016/j.ajogmf.2022.100645.
9. W. Yuan, P. Liu, and G. Neubig, “Can We Automate Scientific Reviewing?,” jair, vol. 75, pp. 171–212, Sep. 2022, doi: 10.1613/jair.1.12862.
10. A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi, “AI-assisted peer review,” Humanit Soc Sci Commun, vol. 8, no. 1, Art. no. 1, Jan. 2021, doi: 10.1057/s41599-020-00703-8.
11. G. R. Latona, M. H. Ribeiro, T. R. Davidson, V. Veselovsky, and R. West, “The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates,” May 03, 2024, arXiv: arXiv:2405.02150. doi: 10.48550/arXiv.2405.02150.
12. A. Saad et al., “Exploring the potential of ChatGPT in the peer review process: An observational study,” Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 18, no. 2, p. 102946, Feb. 2024, doi: 10.1016/j.dsx.2024.102946.

13. M. Thelwall and A. Yaghi, “Evaluating the Predictive Capacity of ChatGPT for Academic Peer Review Outcomes Across Multiple Platforms,” Nov. 14, 2024, arXiv: arXiv:2411.09763. doi: 10.48550/arXiv.2411.09763.

14. G. Kadi and M. A. Aslaner, “Exploring ChatGPT’s abilities in medical article writing and peer review,” Croat Med J, vol. 65, no. 2, pp. 93–100, Apr. 2024, doi: 10.3325/cmj.2024.65.93.

15. G. R. Latona, M. H. Ribeiro, T. R. Davidson, V. Veselovsky, and R. West, “The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates,” May 03, 2024, arXiv: arXiv:2405.02150. doi: 10.48550/arXiv.2405.02150.

16. W. Liang et al., “Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews,” Jun. 15, 2024, arXiv: arXiv:2403.07183. doi: 10.48550/arXiv.2403.07183.

17. J. P. Verharen, “ChatGPT identifies gender disparities in scientific peer review,” eLife, vol. 12, p. RP90230, doi: 10.7554/eLife.90230.

18. Y. Jin et al., “AgentReview: Exploring Peer Review Dynamics with LLM Agents,” Oct. 13, 2024, arXiv: arXiv:2406.12708. doi: 10.48550/arXiv.2406.12708.

19. M. D’Arcy, T. Hope, L. Birnbaum, and D. Downey, “MARG: Multi-Agent Review Generation for Scientific Papers,” Jan. 08, 2024, arXiv: arXiv:2401.04259. doi: 10.48550/arXiv.2401.04259.

20. A. F. Rasheed, M. Zarkoosh, S. F. Abbas, and S. S. Al-Azzawi, “TaskComplexity: A Dataset for Task Complexity Classification with In-Context Learning, FLAN-T5 and GPT-4o Benchmarks,” Sep. 30, 2024, arXiv: arXiv:2409.20189. doi: 10.48550/arXiv.2409.20189.

21. A. Behrouz, P. Zhong, and V. Mirrokni, “Titans: Learning to Memorize at Test Time,” Dec. 31, 2024, arXiv: arXiv:2501.00663. doi:10.48550/arXiv.2501.00663.

22. allenai/marg-reviewer. (Jan. 20, 2025). Python. Ai2. Accessed: Jan. 24, 2025. [Online]. Available: https://github.com/allenai/marg-reviewer

23. J. Cao, “A Study on Deploying Large Language Models as Agents,” Thesis, Massachusetts Institute of Technology, 2024. Accessed: Jan. 22, 2025. [Online]. Available: https://dspace.mit.edu/handle/1721.1/157177

24. J. M. Perkel, “Ten computer codes that transformed science,” Nature, vol. 589, no. 7842, pp. 344–348, Jan. 2021, doi: 10.1038/d41586-021-00075-2.

25. J. M. Perkel, “Why Jupyter is data scientists’ computational notebook of choice,” Nature, vol. 563, no. 7729, pp. 145–146, Nov. 2018, doi: 10.1038/d41586-018-07196-1.

26. A. Vaswani et al., “Attention Is All You Need,” Aug. 02, 2023, arXiv: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

27. “Most Influential ArXiv (Artificial Intelligence) Papers (2024-10),” Paper Digest. Accessed: Jan. 24, 2025. [Online]. Available: https://www.paperdigest.org/2024/10/most-influential-arxiv-artificial-intelligence-papers-2024-10

appendix

Thank you for submitting your paper!
Attention Is All You Need

Your paper will be evaluated by an expert:

Judy Tillman
Expert

Recruiting reviewers...


Marvin Alexander


Robert Brody


Michael Cordova


Lorraine Bishop


Paul Lopez


Floyd Gregg
Editor

High-Level Plan:

1. **Gather Methodology and Experiment Details**: Communicate with other agents to understand the methodology, experimental settings, and key concepts of the paper by sending targeted questions.
2. **Discuss with Expert (Judy Tillman)**: Once I have information on the methodology and experiments, explain them to Expert Judy to solicit her questions and feedback on clarity and reproducibility.
3. **Relay Expert Feedback**: Pass Judy's feedback and questions back to the appropriate agents and gather their responses to improve understanding.
4. **Draft Feedback**: Synthesize the information gathered and Judy's feedback to write a coherent set of clarity and reproducibility-focused feedback comments.
5. **Finalize Review**: Refine the feedback comments for clarity and effectiveness based on responses and my own synthesis. Prepare the final review text.

Step 1: Gather Methodology and Experiment Details

Could you provide a summary of the methodology, experimental settings, and key concepts detailed in the paper? It's important to include how the Transformer architecture within the study is differentiated from traditional methods and the experiments conducted for machine translation and other tasks.

Broadcasting message:

My section focuses on the attention mechanisms, specifically Scaled Dot-Product Attention and Multi-Head Attention, as well as the applications of these mechanisms within the Transformer model. The Transformer uses multi-head attention for encoder-decoder attention, self-attention in the encoder, and self-attention in the decoder to enhance the sequence transduction tasks without using recurrence or convolution. The section notes the values for parameters such as (h = 8) (parallel attention layers), (d_k = d_v = d_{model}/h = 64), and (d_{model} = 512). Additionally, it highlights the position-wise feed-forward networks in the Transformer. For a full summary, I


Robert Brody
Reviewer

position-wise feed-forward networks in the Transformer. For a full summary, I would need details from other sections of the paper regarding the specific experimental settings for machine translation tasks, the description of how the Transformer architecture improves upon traditional methods, and any other key experimental methodologies.

Expected Response: I expect a response detailing the methodology from other sections of the paper, including descriptions of experimental settings, how the Transformer architecture is differentiated from traditional methods, and any specific experiment-related details.

Paul Lopez
Reviewer

The key methodology and experiment details related to my portion of the paper involve evaluating the Transformer model's ability to generalize to English constituency parsing tasks. The methodology included training a 4-layer Transformer model with a model dimension of 1024 on the Penn Treebank dataset (WSJ) with approximately 40K training sentences. Additionally, a semi-supervised approach was used with larger corpora of about 17M sentences, adjusting parameters like vocabulary size, dropout, learning rates, and beam size from a previously established model. This section emphasizes the Transformer model's ability to outperform traditional RNN sequence-to-sequence models, showing the model's efficiency even with limited task-specific tuning.

I expect a response that either confirms my understanding or provides additional details or corrections if necessary, possibly including information about the machine translation experiments or other key concepts mentioned in the abstract that were not covered in my chunk of the paper.

Marvin Alexander
Reviewer

I need more details on the methodology and experiments. Specifically, I require information on the Transformer architecture compared to conventional models, and details about the experimental settings, such as tasks, datasets, and evaluation metrics used. Understanding these will help me explain the approach and results clearly.

Expected Response: The response should include a detailed description of the methodology, covering how the Transformer is designed and differentiated from traditional models, specifics on experimental settings like the tasks employed (e.g., machine translation or others), datasets used, and evaluation metrics or results.

Floyd Gregg
Editor

Step 1 Continued: Gather Methodology and Experiment Details

Based on the responses received, it's clear that each reviewer holds key pieces of the methodology and experimental data, yet there's a need for further integration to fully understand and support Expert Judy Tillman.

Step 2: Discuss with Expert Judy Tillman

Sub-Plan: 1. Provide a summary of gathered information to Judy. 2. Ask her for any specific questions or areas needing clarification based on current understanding.

Expert Judy Tillman, here is a summarization of the methodology and experiment sections I have gathered so far: - Robert highlighted the use of attention mechanisms (Scaled Dot-Product and Multi-Head Attention) within the Transformer model. These methods replace recurrence with self-attention and are crucial in sequence transduction tasks. - Paul detailed the experiment involving English constituency parsing, using a 4-layer Transformer on the Penn Treebank dataset. He noted its superior performance over RNN models and efficiency without extensive tuning. - Marvin requires more on methodological differentiation of Transformer from traditional models, and comprehensive experimental settings across the sections.

Can you provide questions to assess clarity or suggest any improvements regarding the reproducibility or explanation of these methodologies? We are seeking your expertise to refine our understanding and critique.

Expected Response: Judy will provide specific questions about clarity or

concerns about reproducibility that we can relay back to the reviewers for further clarification and exploration.

Broadcasting message: [redacted]

Judy Tillman
Expert

Thank you, Floyd, for providing the summary. Here are some questions and areas where clarity and detail might benefit the paper's methodology and experimental sections:

1. **Attention Mechanisms**:
2. Could you provide more details on how the Scaled Dot-Product Attention and Multi-Head Attention function within this specific application of the Transformer model? Are there any unique modifications or adaptations made to these mechanisms in this study as compared to more conventional settings?
3. **Transformer Model Differentiation**:
4. What specific aspects of the Transformer architecture in the paper differentiate it from traditional methods? It would be useful to have more thorough comparisons or discussions on its theoretical or practical novelty.
5. **Experimental Setup for English Constituency Parsing**:
6. Are there details about the preprocessing of the Penn Treebank dataset and any specific hyperparameters used in the Transformer model training? Having these details is crucial for reproducibility.
9. In terms of maintaining efficiency without extensive tuning, what approaches or default settings were embraced to ensure optimal performance? Some insight into the trial configurations could add much clarity.
10. **General Methodological Detail**:
11. In addition to addressing Marvin's need for methodological differentiation, it would be beneficial to gather specific information on the experimental designs for other tasks beyond English constituency parsing. What benchmarks or datasets were used?

These questions aim to ensure we have comprehensive information on methods and settings to fully understand and reproduce the experiments described in the paper. Please let me know how I can assist further once the information is gathered.

Floyd Gregg
Editor

Step 3: Relay Expert Feedback

Here are questions and areas where more clarity and detail from the paper might benefit us, based on Judy's expertise:

1. **Attention Mechanisms**: Could you provide more details on how Scaled Dot-Product Attention and Multi-Head Attention functionally differ in this study? Are there unique modifications or adaptations made compared to standard applications?
2. **Transformer Model Differentiation**: What specific features differentiate this Transformer model from traditional methods? Judy is particularly interested in the theoretical or practical innovations described in your sections.
3. **Experimental Setup for English Constituency Parsing**: Can you offer additional details on the Penn Treebank preprocessing and specific hyperparameters used during training? Information on metrics used for comparing RNN models would also be helpful.
4. **Efficiency and Tuning**: What was done to maintain efficiency without extensive tuning? Are there details on default or trial configurations that ensured performance, which can be expanded upon?
5. **General Methodological Detail**: Besides addressing the differentiation of Transformer models, do you have insights on experiments and datasets from other tasks beyond English constituency parsing?

Expected Response: Reviewers Robert, Paul, and Marvin should provide detailed explanations or clarifications addressing each of Judy's points regarding the methodologies and experiments they covered.

Broadcasting message: [redacted]

Expected Response: Reviewers Robert, Paul, and Marvin should provide detailed explanations or clarifications addressing each of Judy's points regarding the methodologies and experiments they covered.

Broadcasting message: [redacted]

Robert Brody
Reviewer

1. **Attention Mechanisms**: In my section, the Scaled Dot-Product Attention is described where input queries, keys, and values are processed together as matrices (Q, K, V) and the dot products of queries with keys are scaled by (1/\sqrt{d_k}). Multi-Head Attention extends this by projecting the queries, keys, and values into multiple subspaces, performing attention in parallel for each, allowing the model to capture different aspects of the input through these different attention heads.
2. **Transformer Model Differentiation**: The key differentiator for the Transformer in my section is its reliance on these attention mechanisms, particularly removing sequence operations like recurrence and convolution entirely, which is a departure from traditional neural network models. The architecture's innovation lies in leveraging self-attention for both encoders and decoders, which improves parallelization and reduces training times significantly while maintaining or improving performance.

For other questions, I would need input from Paul and Marvin on details regarding English constituency parsing and general methodological applications beyond my attention mechanism focus.

Expected Response: I expect Paul to provide more details on the experimental setup, particularly for the English constituency parsing task, and Marvin to address questions about the broader methodological applications and setups from other parts of the study.

Marvin Alexander
Reviewer

1. For the attention mechanisms, could someone elaborate on how Scaled Dot-Product Attention and Multi-Head Attention are used in the Transformer, and whether there are any unique modifications or adaptations compared to standard applications?
2. Regarding the Transformer model differentiation, I need insights into what specifically sets this model apart from traditional methods. Are there particular theoretical or practical innovations highlighted?
3. Can someone else elaborate on the experimental setup for English constituency parsing, especially on Penn Treebank preprocessing, hyperparameters used during training, and metrics for comparing RNN models?
4. What strategies are employed to ensure efficiency in the model without extensive tuning? Are there details about default settings or trial configurations for performance?
5. I need more information about experiments and datasets related to tasks beyond English constituency parsing. Details on different tasks and their assessments would be useful.

Expected Response: The response should include detailed explanations for each of the points above, addressing Judy's request for clarity and providing comprehensive understanding aligned with the feedback needs.

Paul Lopez
Reviewer

1. In my section of the paper, specific details about Scaled Dot-Product Attention or Multi-Head Attention are not provided, so I can't offer additional functional differences or unique adaptations from standard applications. I might need to coordinate with the reviewer covering that section for comprehensive information.
2. The differentiation of the Transformer model in terms of features from traditional methods is not explicitly detailed within my paper chunk, which primarily covers its application to English constituency parsing. However, the abstract mentions no reliance on recurrence and convolution. I require further details from sections covering traditional methods for a theoretical or practical innovation comparison.

3. Regarding the experimental setup, we used the Wall Street Journal (WSJ) section of the Penn Treebank. The paper mentions using a 4-layer transformer model with a model dimension (d model) of 1024, a vocabulary size of 16K for WSJ-only, and 32K tokens for semi-supervised settings. Dropout, learning rates, and beam size were adjusted using the Section 22 development set, with a beam size of 21 and α = 0.3. Metrics were compared to the Berkeley-Parser and other RNN models, but I need more specifics about preprocessing or precise hyperparameters if they're detailed elsewhere.

4. Efficiency was mentioned relative to empirical improvements observed with minimal task-specific tuning. The reliance on existing parameters from the English-to-German translation model denotes a preference for stability and avoiding unnecessary tuning experimentation. More insights may be required from other sections if specific configurations ensured performance.

5. My paper chunk does not cover tasks beyond English constituency parsing. For insights on other experiments or datasets, it would be beneficial to follow up with reviewers covering sections on machine translation or other tasks.

I expect responses from the other reviewers, especially Robert and Marvin, providing details or directing me to where in the paper these elements are clarified. If they need information from my section to enrich their answers, I'll be ready to assist.

Floyd Gregg
Editor

Step 4: Draft Feedback

Based on the responses received from Expert Judy Tillman and review team members, it is clear there is a need to connect the details across all reviewed sections of the paper to fully address the feedback queries. I will synthesize the responses to provide a coherent set of comments focusing on Clarity and Reproducibility.

Draft Feedback Comments:

1. Attention Mechanisms: - **Clarity**: Robert provided a detailed explanation of the Scaled Dot-Product Attention and Multi-Head Attention within the Transformer model, highlighting its implementations across encoders and decoders without requiring sequence operations like recurrence. To enhance clarity, it would be beneficial for each reviewer to coordinate to ensure unique modifications, if any, are clearly stated as compared to standard applications.

2. Transformer Model Differentiation: - **Reproducibility & Innovation**: While Robert covers the innovative use of self-attention, a comprehensive summary including practical comparisons with conventional methods, as suggested by Marvin, can strengthen the theoretical novelty section. Cross-referencing parts of the paper detailing transitions from traditional RNN or convolution-based models could set more context as Paul suggested.

3. Experimental Setup for English Constituency Parsing: - **Detail for Reproducibility**: Paul provided an overview of experimental variables for English constituency parsing but needs to cross-collaborate with Robert and potentially other sections for specifics on preprocessing steps and hyperparameter selections. Details on metrics used would solidify the comparative ground against RNN models.

4. Efficiency and Tuning: - **Clarity on Approaches**: The reliance on previously established parameters underscores efficiency, but detailing any key trial configurations, especially in unparalleled sections, would give readers practical insights into the minimal tuning required.

5. General Methodological Detail: - **Comprehensiveness**: As indicated by Marvin, there needs to be a cohesive approach to covering experimental tasks beyond constituency parsing. Connections to datasets and experiments from machine translation, effectively synthesized, would present a full spectrum of applicability and strengthen the paper’s methodology section.

Step 5: Finalize Review

Finalizing: I will ensure comments align well with Judy's queries and the detailed responses from my team.

Final Version of the Feedback Review:

This feedback prioritizes clarity in presenting attention mechanisms, reproducibility through improved detailing of experimental setups like English constituency parsing, and drawing clearer contrasts between the novel Transformer model and traditional neural networks. Each section can bolster its contribution by ensuring cross-references amongst extensions and guarantees to collectively narrate a comprehensive methodology.

Integration Feedback Recommendations:

- Encourage cross-section clarity: Researchers must ensure unique adaptations or innovations are distinctly noted without reader assumptions of standard applications.
- Amplify reproducible detailing: Ensure that experimental setups are richly detailed in preprocessing and parameterization, especially for setups not covered in their entirety within one section.
- Deepen innovation narrative: Align differences from traditional methods consistently across the paper to reinforce the innovation narrative.

Conclusion:

By incorporating these feedback areas, the paper will not only richly describe the functional and theoretical significance of the Transformer architecture but also lay down a robust blueprint for reproducibility that other researchers can confidently follow.

READY