November 2021

# Efficient NLP

Yuki Arase, Osaka University, arase@ist.osaka-u.ac.jp
Phil Blunsom, Oxford University & Deepmind, pblunsom@google.com
Mona Diab, George Washington University & Facebook AI, mtdiab@gwu.edu
Jesse Dodge, Allen Institute for AI, jessed@allenai.org
Iryna Gurevych, Technical University of Darmstadt, gurevych@ukp.informatik.tu-darmstadt.de
Percy Liang, Stanford University, pliang@cs.stanford.edu
Colin Raffel, UNC & HuggingFace, craffel@gmail.com
Andreas Rücklé, Amazon, andreas@rueckle.net
Roy Schwartz, The Hebrew University of Jerusalem, roy.schwartz1@mail.huji.ac.il
Noah A. Smith, University of Washington & Allen Institute for AI, nasmith@cs.washington.edu
Emma Strubell, Carnegie Mellon University & Google, strubell@cmu.edu
Yue Zhang, Westlake University, yue.zhang@wias.org.cn

The amount of computation put into training NLP models has grown tremendously in recent years. This trend raises the bar for participation in NLP research, excluding large parts of the community from experimenting with state-of-the-art models. It also creates environmental concerns since this computation uses increasing amounts of energy. This document is the report of a working group appointed by the ACL Executive Committee to promote ways that the ACL community can reduce the computational costs of NLP and thereby mitigate some of these concerns. The recommendations in this report are guided in part by the results of a survey we conducted with the ACL community[1] as well as a presentation of the main ideas during the ACL 2021 business meeting.[2] Below we provide a summary of our recommendations, followed by a background section, a detailed description of our recommendations, and a list of additional ideas proposed by members of the community.

## *1. Summary of recommendations*

(i) Increasing the alignment between *experiments* and *research hypotheses*. The amount of computational resources a researcher can afford does not dictate the novelty of their ideas and the scientific validity of their arguments. We suggest putting more emphasis on the scientific rigour of papers in our community. To achieve this goal, we aim to release more thorough guidelines for authors and reviewers. *Authors* need to justify their experimental setups. Likewise, *reviewers* should critique the experimental setup and results in terms of their support for a paper's claims, rather than their position on some leaderboard. We suggest adding a question to the review form asking to which extent the experiments in the paper support the research hypotheses.
(ii) Encouraging the release of trained models*, potentially of varying sizes, along with the training code,* as a mechanism for reducing the computation cost necessary to build upon or reproduce scientific results, as well as saving the human time required to reimplement them. For example, we suggest rewarding the release of pretrained models or training sets through the review process, and/or by making a visual branding in the conference website of papers that release models, code, and training data.

[1]For a summary of the survey results, see https://forms.office.com/Pages/AnalysisPage.aspx?id=9028kaqAQ0OMdrEjlJf7WQiNRJRoOx9OlzQS6C5hck5UQkQ2S0s3UzdYTjU4NkhRODkwMDhWMUQ0SC4u&AnalyzerToken=KXeMl5e44GizvOxam8drb8BLE5SbKTV1
[2]https://underline.io/events/167/sessions/6380/lecture/31194-green-nlp-panel

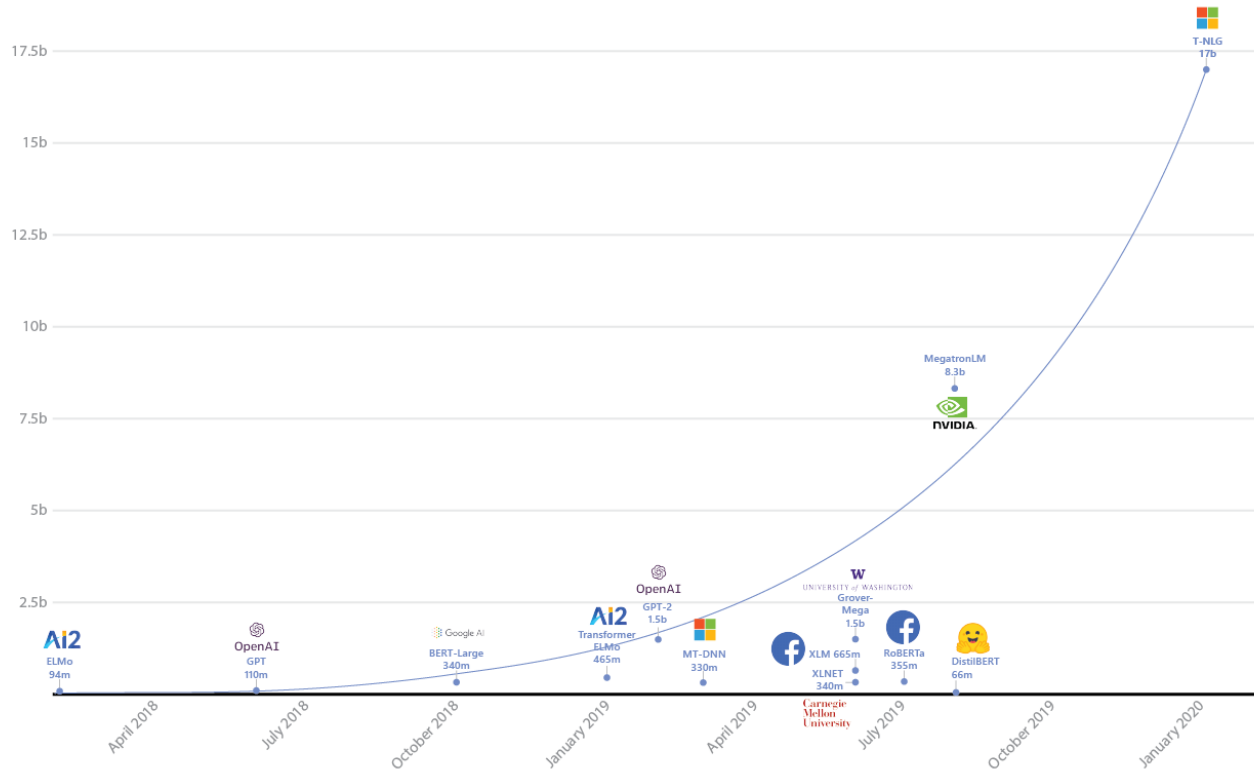


Figure 1: The growth in the number of parameters of NLP models, starting from ELMO until T-NLG.

(iii) *Setting up conference tracks that target efficiency.* A track that targets efficiency directly will promote work that aims to reduce the cost of NLP experimentation and deployment, and match submissions to qualified reviewers that appreciate the nuances of evaluating this type of work. Some specific types of submissions could include work on model compression, sample efficiency, early stopping, and research showing that simple and efficient approaches outperform more complex ones.

### *2. Background and motivation*

In the past decade, the field of AI, and NLP in particular, has reported remarkable progress on a broad range of capabilities, including question answering, natural language inference, and machine translation. Much of this progress has been achieved by increasingly large and computationally intensive deep learning models. Figure 1[3] plots the increase in number of parameters over time for NLP word-embedding approaches, starting with ELMo [1] in late 2017 up to Turing-NLG [2] in early 2020. The trend shows an increase in parameter count of 175x in little over 2 years. This trend further continued into 2020 with a growth of more than 50x, following the release of GPT-3 [3] and Switch Transformer [4] (a total of 10,000x in 3 years). Recent works [5,6] have estimated the carbon footprint of several NLP models and argued this trend is both environmentally unfriendly and prohibitively expensive, raising barriers to participation in NLP research.

The results of the survey we conducted with the ACL community reveal that this is a real concern in our community: 57% of the 312 responders report that in the past year, they were unable to run experiments

[3] Reproduced from https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

that are important for one of their projects due to computational constraints. Further, 70% report that they sometimes feel that their work would have been valued more by the community had they had access to more computational resources. Indeed, 30% of the responders have received reviews that requested experiments too expensive for their budget for a particular paper. Finally, 75% of the responders are at least somewhat concerned by the environmental footprint of the field of NLP.

To address these concerns, we recommend putting more thought into the conditions in which an expensive experiment is required, by increasing the alignment between *experiments* and *research hypotheses*. Our goal is for both authors and reviewers to justify the link between the experiments run (and those that were not run) and the research questions in the relevant paper. We would like to encourage researchers to think about whether certain experiments are necessary or not, which will then lead to lower energy cost, to a more inclusive environment, and to increased scientific rigour.

We also propose to take additional measures to promote efficiency, by both encouraging the release of trained models, which would save compute for others, and the introduction of a dedicated track for efficient methods for NLP that would promote more work on that topic. Below we describe each of our proposals in detail.

### *3. Recommendations*

*(a) Increasing the alignment between experiments and research hypotheses.*

A good scientific paper typically states clear research hypotheses and describes results (experimental, theoretical or other) that support these hypotheses. Designing the experimental setup that directly addresses the corresponding hypotheses is one of the basic skills any researcher acquires during their research training. The role of a scientific paper is, in part, to express the logic behind the choice of experiments, and to convince the reader that the results indeed support the hypotheses. A common pitfall in modern NLP is to mistake quantity with quality, and thus over-rely on the computational budget when evaluating a paper. Our goal is to increase the scientific rigour of NLP papers, by making a stronger connection between experiments and research hypotheses. We explicitly want to reduce the incentive for leaderboard/SOTA chasing, which naturally leads to high computational costs. Instead, we aim to ensure that the experiments carried out are sufficient to effectively prove the research hypotheses and unnecessary experiments are avoided.

The amount of computational resources that a researcher has access to is certainly related to their ability to conduct good research,[4] but is far from sufficient. For instance, a paper that introduces a new feature for a certain architecture might experiment with a small variant of a popular pretrained model. If this author conducted well-planned experiments, comparing against all the relevant baselines, it is likely that their experimental setup indeed supports their research hypothesis, and it is thus scientifically valid. In contrast, a low computational budget cannot justify a poor experimental setup, which misses important baselines or ablations. Similarly, a large computational budget can be used in different ways. Running multiple experiments can be done in a principled and rigorous way, and lead to valuable scientific

[4]E.g., at the extreme, a researcher with no computational budget whatsoever will not be able to run any experiments.

findings. At the same time, a large budget doesn't guarantee that the right comparisons were conducted by the authors, and that the scientific claims are supported.

Many papers published in NLP today draw conclusions based on experimental results from models with a few hundred million parameters, around the size of BERT-Large, and some reviewers might consider experiments with smaller models insufficient; this is due to the implicit doubt that conclusions from these experiments might not generalize to a larger regime. However, whether results generalize to larger models is its own research question, which warrants its own evaluation, and this more expensive evaluation may simply be out of scope for some projects. The range of "appropriate" model sizes is continually changing, and many results presented at NLP conferences today would not be expected to generalize to the largest models (which can have trillions of parameters). In fact, some researchers intend to focus only on smaller models, like those designed to fit on phones or those with low latency for interacting with people, and thus aren't trying to address generalization to larger models in their work.

We suggest releasing better guidelines for authors and reviewers that target this issue. Importantly, these guidelines should address both authors and reviewers. *Authors* are requested to justify their experimental setups. What are the research hypotheses? How are the experiments supporting these hypotheses? What alternative explanations exist for the presented results, and how do the authors rule them out? They should also set the reader's expectation by carefully articulating their computational budget, along with justifications for why that is appropriate to address their stated research hypotheses. Given that these justifications are valid, reviewers should support research on smaller models without expecting larger-budget evaluations, as such research can still provide valuable insights. *Reviewers* in turn should use scientific terms to justify their reviews. They should be mindful to the extent to which the experiments support the research hypotheses. We suggest adding this question explicitly to the review form. Further, reviewers are allowed (and encouraged!) to ask for additional experiments that would better support the research hypotheses, but they should clearly state in turn *their hypotheses* as to the role of the new experiments.

*(b) Encouraging the release of trained models*.

Releasing and sharing trained models is one of the most effective ways of reducing computational costs: if someone already ran some code, there is no reason to run it again on the same inputs. Our goal is to provide incentives for researchers to release their trained models. A key aspect of this effort is to not only release the largest, most expensive models, but also smaller models. This is important because some models are too expensive to even load into memory for some authors (and in some hardware settings), so releasing smaller models would provide these authors with a way to run experiments that fit their budget, and in general provide all authors (even those who can and wish to work with larger models) with smaller and cheaper models to run their preliminary experiments on. Further, we also recommend releasing the output of the main experiments (e.g., the model predictions on the relevant datasets) to foster cheaper and faster analysis of the provided models. Finally, to promote reproducibility and faster research cycles, we also want to encourage authors to release their training code.

In order to implement this principle, we proposed several concrete recommendations, and asked the survey participants to endorse the ones they support. Below we provide the two most well-received recommendations, in decreasing order of support.

- *Visible branding of papers that release models/code*. Such papers could be marked with an asterix or some other badge (similar to ACM's artifact review badging) at the ACL anthology website, including an easy access to the code link (similarly to the PDF and citation links). They can also be marked on the particular conference website as such. This proposal received support from 35% of the survey responders.
- *Instructing reviewers to reward papers who share their code*. This point, supported by 30% of the responders, raises a few delicate issues. First, it is not clear what is considered to be code sharing. Some submissions upload their code/data during submissions, others promise to do so, some upload an incomplete/poorly documented repository, and so on. Further, it is not clear that looking into other people's code, or even just running it, is the best use of reviewers' time. The input we got from the community indicates that it is important to devise policies around this issue. We therefore make the following recommendations:
  - To incentivize code and model release, the deadline for releasing the code will be one week after the conference deadline, which would allow authors to clean up / package the code in a time that does not conflict with writing the paper. This will encourage authors to release their code and data, and to put more effort into making the code user-friendly. This is similar in spirit to the NeurIPS policy for supplementary materials. Authors who release code will be encouraged to follow code writing policies (e.g., the NeurIPS guidelines).
  - Importantly, code/data release would still be voluntary, and a submission without these is still a valid submission. Authors can choose whether to upload their code along with their submission, to promise to do so, or not to address this point at all.
  - Reviewers, similar to the current practice, are the ones in charge of evaluating the code as they see fit. Importantly, reviewers are **not** expected to download and/or run the code (although they are more than welcome to do so), but merely to check for inconsistencies (e.g., a paper claiming to release code, which doesn't deliver).
  - Reviewers will be encouraged to reward papers based on the paper's chosen policy. I.e., reward those who share both code and models most, followed by those who share just one of them, those who promise to share them, and finally, those who neither share nor promise to share will not be rewarded at all in this section. Given that every paper is different and we would like to still provide reviewers with the freedom to assess the papers as they see fit, we do not provide concrete recommendations on how much to reward these papers, but rather a general guideline.
  - Further, we do **not** instruct reviewers to **penalize** papers that don't share their code and models, as we recognize that this is not always possible due to privacy concerns or other regulatory requirements. Such papers could consider submitting to industry tracks dedicated for publishing research on models/data that cannot be shared.
  - Finally, we do not recommend a monitoring process to verify that authors that promise to release models or code actually do so. Our community is built on trust, and similarly to how we do not verify that authors take reviewers recommendation into account, we believe that authors promising to release models or code do so in good faith.

(c) *Setting up tracks that target efficiency.*

Works in *ACL conferences often focus on improving accuracy rather than efficiency measures such as runtime and number of parameters [6]. As a result, some authors might feel dis-incentivized to pursue research questions that target efficiency gains. We believe such work is valuable and could help mitigate some of the environmental and social concerns raised in this document. To promote such work, and inspired by efforts in other AI communities,[5] we propose setting a dedicated track for work on efficiency. Examples of typical submissions to this track will include:

1. Models that train on smaller amounts of data (sample efficiency)
2. Models with a smaller number of parameters (memory efficiency)
3. Models that require fewer computational resources, during training and/or inference
4. Efficient model selection methods (e.g., efficient hyperparameter search)
5. Methods for better reporting of computational budgets (e.g., more accurate methods for measuring carbon emissions)
6. Methods for extrapolating results on smaller models to larger ones

Note that this track is ***not*** meant for submissions which use NLP for positive impact on the environment, e.g., to mitigate climate change, but rather on efficient solutions for NLP. Moreover, like other general tracks, such as machine learning for NLP, this track is not dedicated to a particular application, but rather focuses on the methods for making models more efficient.

Our survey indicates that almost 90% of responders working on efficient methods would consider submitting to such a track. In the past year, a similar track was run at EACL 2021, NAACL 2021, and EMNLP 2021. The adoption of these tracks was surprisingly high: at EMNLP 2021, over 130 papers were submitted to the Efficient NLP track. Further, members of this working group have served as area chairs and senior area chairs of all these tracks. Our experience was that recruiting reviewers was relatively simple and that the level of papers was not substantially different from the general *ACL pool. We suggest making this track a permanent track in all *ACL conferences.

Finally, this proposal assumes the current structure of *ACL conferences, which have topical tracks. The new ACL rolling review works differently, and does not include tracks, which might impact some of our recommendations. Nonetheless, assuming the main conferences–who set the calls for papers and make acceptance decisions–still keep the track structure, much of the value of our suggestions still holds.

### *4. Other ideas*

Apart from the three concrete suggestions above, below we describe additional suggestions that were either not supported by the survey participants, or we believe are too complex to implement at this point. We lay them out here, as they might be adopted in the future.

1. Establishing a “Best Artifact Award” (as in, e.g., OOPSLA) for papers that release the pretrained model along with the data, code, and processes for producing the model.

[5] E.g., CVPR 2020, ICCV 2021.

2. Raising the publication bar for papers that release huge models (e.g., stricter code release requirement, stronger justification of the computational budget).
3. Require papers to disclose details about the amount of computational resources invested in the paper (including preliminary experiments), as well as the amount of carbon emitted in this process.
4. Encourage virtual conferences in order to reduce the carbon footprint of travel to remote conferences.
5. Pre-registration of experiments to avoid the duplication of work (see [7]).
6. Introduce emission caps on NLP papers, or at least require offset purchases for papers with high emissions.

***References***


[1] Peters et al., *Deep contextualized word representations*. In Proc. of NAACL 2018

[2] *Turing-NLG: A 17-billion-parameter language model by Microsoft*. Microsoft Research Blog. 2020.

[3] Brown et al., *Language Models are Few-Shot Learners*. In Proc. of NeurIPS 2020.

[4] Fedus et al., *Switch Transformers: Scaling to Trillion Parameter Models*. arXiv:2101.03961

[5] Strubell et al., *Energy and Policy Considerations for Deep Learning in NLP*. In Proc. of ACL 2019

[6] Schwartz et al., *Green AI*. In CACM. 2020.

[7] van Miltenbur et al., *Preregistering NLP Research*. In Proc. of NAACL 2021