Reproducible Machine Learning

Originalartikel

Backup

<html> <div class=„“><h2 id=„5fec“ class=„hs gt gu bf b ht hu hv hw hx hy hz ia ib ic id ie if ig ih ii dx“>A step towards making ML research open and accessible</h2><div class=„db n cp ij ik il“><div class=„o n“><div><a href=„https://medium.com/@preetihemant?source=post_page-----cf1841606805———————————————–“ rel=„noopener follow“><img alt=„Preeti Hemant“ class=„s im in io“ src=„https://miro.medium.com/fit/c/56/56/2*NekPqscrKMBfu8yd2bygng.png“ width=„28“ height=„28“ referrerpolicy=„no-referrer“ /></a></div><div class=„ip v n cy“><div class=„n c1“><a href=„https://medium.com/@preetihemant?source=post_page-----cf1841606805———————————————–“ class=„“ rel=„noopener follow“><p class=„bf b bg bh fo“>Preeti Hemant</p></a></div><a class=„“ rel=„noopener follow“ href=„https://towardsdatascience.com/reproducible-machine-learning-cf1841606805?source=post_page-----cf1841606805———————————————–“><p class=„bf b bg bh dx“>Feb 17, 2020&#183;4 min read</p></a></div></div></div></div><figure class=„jh ji ga gb paragraph-image“><div role=„button“ tabindex=„0“ class=„jj jk at jl v jm ga gb jg“><img alt=„“ class=„v jn jo“ src=„https://miro.medium.com/max/1400/1*IL2eR2U9D6cttbyioQw-8Q.jpeg“ width=„700“ height=„233“ role=„presentation“ referrerpolicy=„no-referrer“ /></div><figcaption class=„jp jq gc ga gb jr js bf b bg bh dx“>Photo credit: <a href=„https://pixabay.com/users/geralt-9301/“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>geralt</a> via <a href=„https://pixabay.com“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>Pixabay</a></figcaption></figure><p id=„9e93“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>The <a href=„https://nips.cc“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>NeurIPS</a> (Neural Information Processing Systems) 2019 conference marked the third year of their annual reproducibility challenge and the first time with a reproducibility chair in their program committee.</p><p id=„151e“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><em class=„kq“>So, what is reproducibility in</em> <a href=„https://www.mathworks.com/discovery/machine-learning.html“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“><em class=„kq“>machine learning</em></a><em class=„kq“>?</em></p><p id=„08a5“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><strong class=„jw gv“>Reproducibility</strong> is the ability to be recreated or copied. In machine learning, reproducibility is being able to recreate a machine learning workflow to reach the <strong class=„jw gv“>same conclusions</strong> as the original work.</p><p id=„747b“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><em class=„kq“>Why is this important?</em></p><p id=„e037“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>An algorithm from new research without the reproducibility aspects can be difficult to investigate and implement. With our increasing dependency on ML and AI systems for decision making, integrating a model not fully understood can have unintended consequences.</p><p id=„eb04“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>Costs and budget constraints are another area impacted by reproducibility. Without the details on hardware, computational power, training data and more nuanced aspects like hyper-parameters tuning, adopting new algorithms can run into huge costs and considerable research effort, only to lead into inconclusive results.</p><figure class=„ks kt ku kv kw ji ga gb paragraph-image“><figcaption class=„jp jq gc ga gb jr js bf b bg bh dx“>Reproducibility, a factor in building trustworthy ML (Photo by author)</figcaption></figure><p id=„a8e7“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><em class=„kq“>How to

identify if a ML model/research meets reproducibility standards?</em></p><p id=„63d5“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>This post explores two approaches proposed in research papers to establish reproducibility:</p><p id=„13da“ class=„ju jv gu jw b ht lx jy jz hw ly kb kc kd lz kf kg kh ma kj kk kl mb kn ko kp gn hr“>The authors suggest missing information as the root cause of all reproducibility issues. Missing information could be intentional (trade secret) or unintentional (undisclosed assumption). They focus on methods to rectify the case of unintentional problems that hinder reproducibility.</p><p id=„60a2“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>Aspects impacting reproducibility and the solutions suggested are summarized here:</p><p id=„8828“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><a href=„https://hackernoon.com/what-is-training-data-really-adf0b97a116c“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“><strong class=„jw gv“>Training data</strong></a><strong class=„jw gv“>:</strong> A model produces different results for different training sets. One way of addressing this is to adopt a versioning system that records changes in the training data. However, this may not be practical for large data sets. Using data with documented timestamps is a workaround. The other option is to save hashes of data at regular intervals with documentation of the calculation methods.</p><p id=„c5f4“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><a href=„https://www.datarobot.com/wiki/feature/“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“><strong class=„jw gv“>Features</strong></a> <strong class=„jw gv“>:</strong> Features can produce varying results based on how they are selected and generated. The steps taken in generating features should be tracked and version controlled. Other best practices include i) keeping individual feature generation code independent from one another ii) in case of a bug fix for a feature, creating a new feature.</p><p id=„fcb5“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><a href=„https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“><strong class=„jw gv“>Model training</strong></a><strong class=„jw gv“>:</strong> Documenting details of how the model was trained will ensure repeatable results. Feature transformations, order of features, hyperparameters and the method to select them, structure of the ensemble are among important model training details to be maintained.</p><p id=„180a“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“><strong class=„jw gv“>Software environment:</strong> Software versions and packages used also play a role in replicating results seen in the original ML model. An exact match in software used may be required. Hence, even if there have been software updates, it is best to use the version on which the model was originally trained.</p><p id=„b396“ class=„ju jv gu jw b ht lx jy jz hw ly kb kc kd lz kf kg kh ma kj kk kl mb kn ko kp gn hr“>Here, the authors attempt to independently reproduce (without use of the original author&#8217;s code) two hundred and fifty five papers. They record attributes of the research papers and analyze the correlation between reproducibility and the attributes using <a href=„https://en.wikipedia.org/wiki/Statistical_hypothesis_testing“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>statistical hypothesis testing</a> .</p><p id=„9cc6“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>They use a total of 26 attributes per paper.</p><p id=„c60c“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>Unambiguous (require no explanation): <em class=„kq“>Number of Authors, the existence of an appendix (or supplementary material), the number of pages (including references, excluding any appendix), the number of references, the year the paper was published, the year first attempted to implement, the venue type (Book, Journal, Conference, Workshop, Tech-Report), as well as the specific publication venue (e.g., NeurIPS, ICML).</em></p><p id=„2dec“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>Mildly subjective : <em class=„kq“>Number of Tables,Number of Graphs/Plots, Number of Equations, Number of Proofs, Exact Compute Specified, Hyper-parameters Specified, Compute Needed, Data Available, PseudoCode.</em></p><p id=„d63e“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>Subjective : <em class=„kq“>Number of Conceptualization Figures, Uses Exemplar Toy Problem, Number of Other Figures, Rigor vs Empirical, Paper Readability, Algorithm Difficulty, Primary Topic, Looks Intimidating.</em></p><h2 id=„20e7“ class=„lh li gu bf lj lk ll hv lm ln lo hy lp hz lq ib lr ic ls ie lt if lu ih lv lw hr“>Results and Significant relationship</h2><p id=„bcf5“ class=„ju jv gu jw b ht lx jy jz hw ly kb kc kd lz kf kg kh ma kj kk kl mb kn ko kp gn hr“>To determine significance, authors use non-parametric <a href=„http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric4.html“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>Mann&#8211;Whitney U test</a> for numeric features and <a href=„https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>Chi-squared test</a> with <a href=„https://www.statisticshowto.datasciencecentral.com/what-is-the-continuity-correction-factor/“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>continuity correction</a> for categorical features.</p><p id=„2811“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>Out of the twenty six attributes, ten are significantly correlated to reproducibility.</p><p id=„a78c“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>These are <em class=„kq“>Number of Tables, Equations, Compute Needed, Pseudo-Code, Hyper-parameters Specified, Readability, Rigor vs Empirical, Algorithm Difficulty, Primary Topic, Authors Reply.</em></p><p id=„5ced“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>The more information there is about the significant attributes, easier it is to reproduce a paper. However, <em class=„kq“>Number of equations</em> is negatively correlated to reproducibility. This does seem counter-intuitive and the authors offer two theories as to why: <em class=„kq“>1) having a larger number of equations makes the paper more difficult to read, hence more difficult to reproduce or 2) papers with more equations correspond to more complex and difficult algorithms, naturally being more difficult to reproduce.</em></p><h2 id=„056f“ class=„lh li gu bf lj lk ll hv lm ln lo hy lp hz lq ib lr ic ls ie lt if lu ih lv lw hr“>Study limitations</h2><p id=„4498“ class=„ju jv gu jw b ht lx jy jz hw ly kb kc kd lz kf kg kh ma kj kk kl mb kn ko kp gn hr“>Authors in this study select papers based on personal interests. The topics are not randomly picked, hence a <a href=„https://en.wikipedia.org/wiki/Selection_bias“ class=„ek jt“ target=„_blank“ rel=„noopener ugc nofollow“>selection bias</a> is introduced. Also, the results are to be taken with the consideration that the authors are not experts on the topics they select.</p><p id=„75dd“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>Thirdly, there are many subjective attributes that are significant, and need developing objective measures related to the subjective factors. Eg &#8212; measuring paper readability or algorithm difficulty.</p><p id=„bf49“ class=„ju jv gu jw b ht lx jy jz hw ly kb kc kd lz kf kg kh ma kj kk kl mb kn ko kp gn hr“>The first study takes a bottom-up approach while the second a top-down. Both together give an insight into the factors necessary for replicating machine learning pipelines and reproducing original results. As seen from the results of the two approaches, model hyper-parameters and computing resources/software environment are two important factors influencing reproducibility.</p><p id=„1b42“ class=„ju jv gu jw b ht jx jy jz hw ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp gn hr“>In general, it is important to record every step taken in building a model through documentation and version control. After all, many machine learning systems are black boxes and considering reproducibility makes the research open, accessible and reproducible.</p> </html>