TOWARDS REPRODUCIBLE ML RESEARCH IN NLP

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Robert Stojnic

ACL 2022

MECHANISMS FOR REPRODUCIBILITY

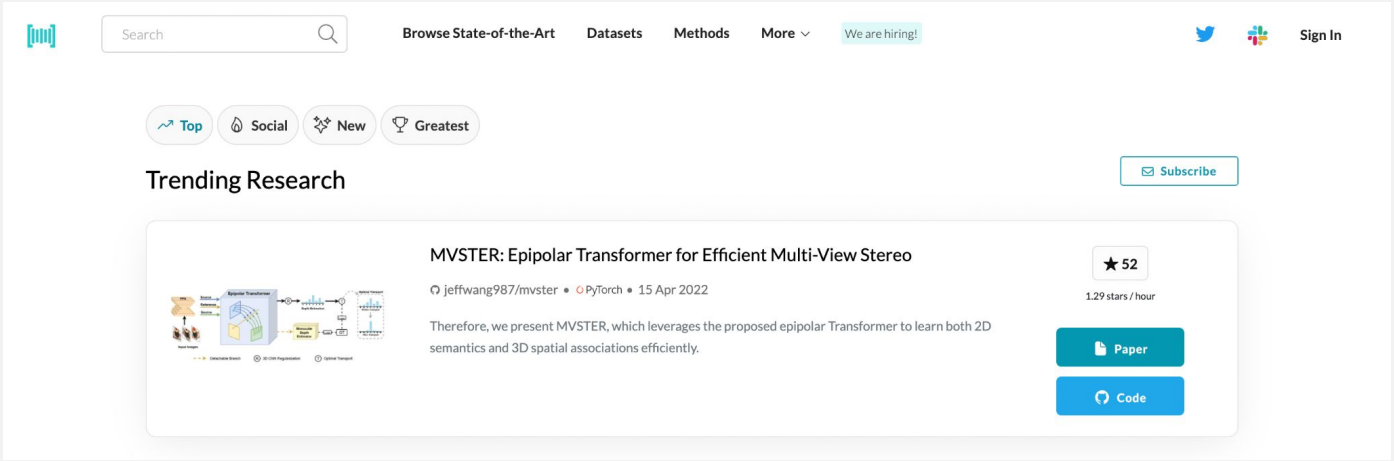
Koustuv Sinha, Robert Stojnic, Jessica Zosa Forde

OVERVIEW

1. Papers with Code
2. Reproducibility Challenge
3. Reproducibility Checklists
4. Useful Tools and libraries

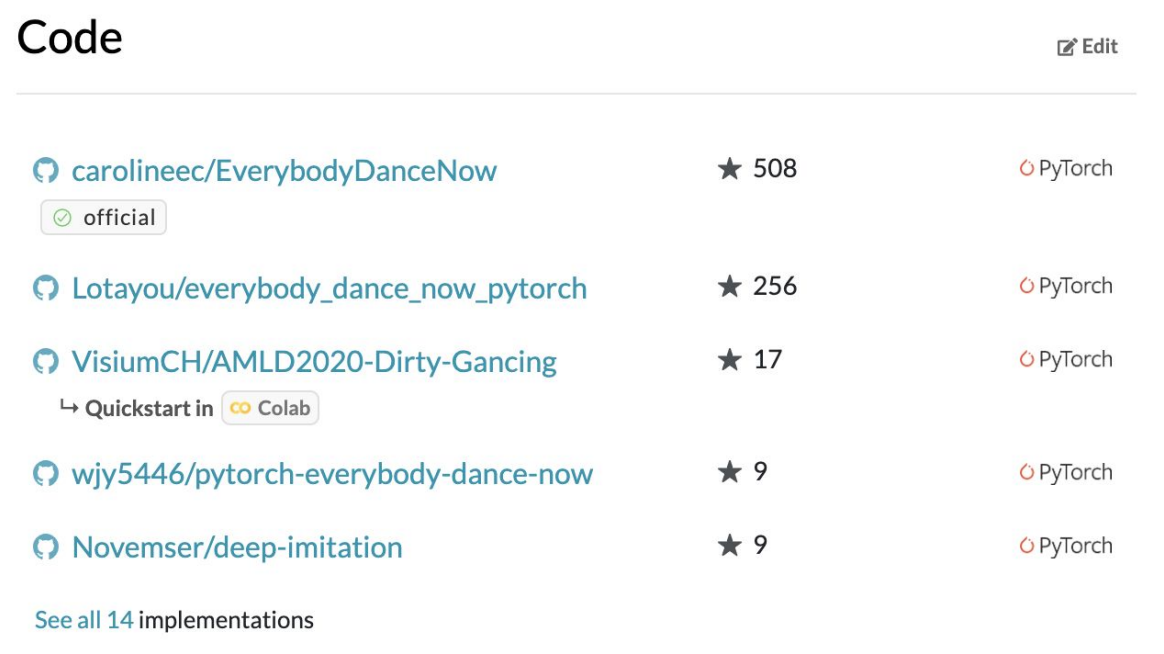
PAPERS WITH CODE

- **Goal**: Track all artefacts in ML, create positive incentives for sharing



PAPERS WITH CODE

- Largest database of papers curated with their code



PAPERS WITH CODE

- Largest database of datasets, tracking their usage

ImageNet

Edit

Introduced by Jia Deng et al. in ImageNet: A large-scale hierarchical image database

The **ImageNet** dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection. The publicly released dataset contains a set of manually annotated training images. A set of test images is also released, with the manual annotations withheld. ILSVRC annotations fall into one of two categories: (1) image-level annotation of a binary label for the presence or absence of an object class in the image, e.g., "there are cars in this image" but "there are no tigers," and (2) object-level annotation of a tight bounding box and class label around an object instance in the image, e.g., "there is a screwdriver centered at position (20,25) with width of 50 pixels and height of 30 pixels". The ImageNet project does not own the copyright of the images, therefore only thumbnails and URLs of images are provided.

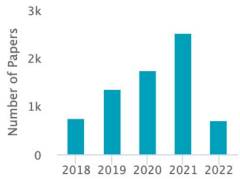
- Total number of non-empty WordNet synsets: 21841
- Total number of images: 14197122
- Number of images with bounding box annotations: 1,034,908
- Number of synsets with SIFT features: 1000
- Number of images with SIFT features: 1.2 million

Source: ImageNet Large Scale Visual Recognition Challenge

Homepage

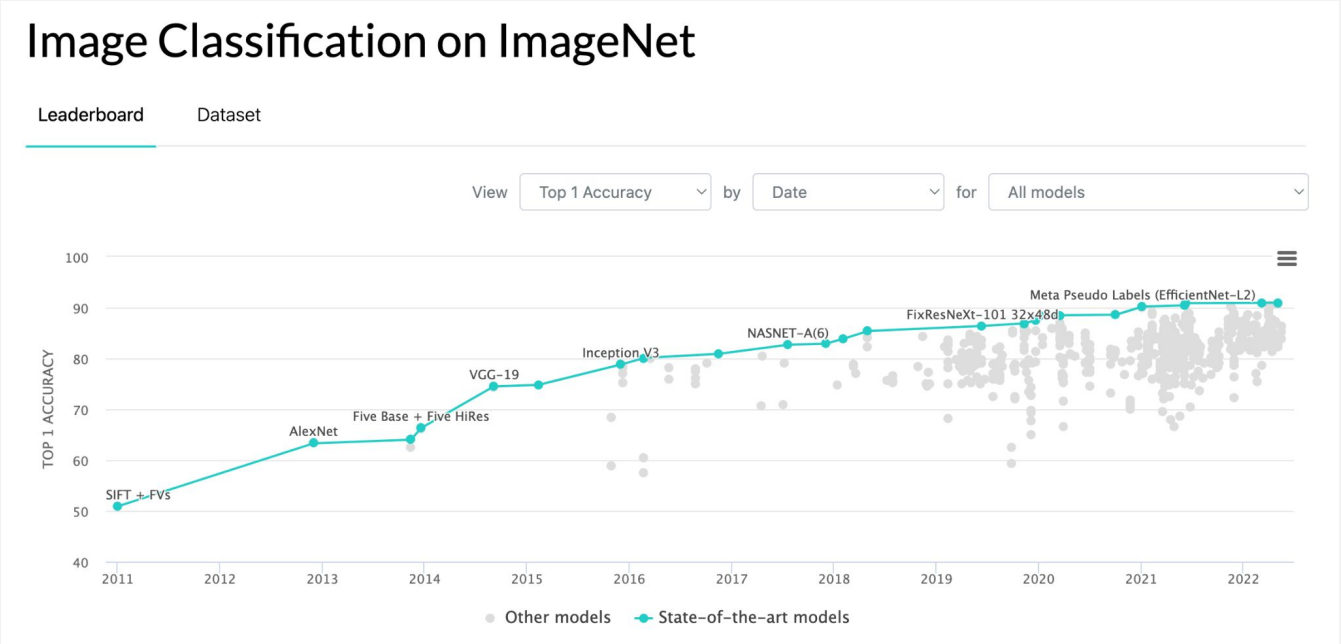
Source: https://cs.stanford.edu/people/kar...

Usage



PAPERS WITH CODE

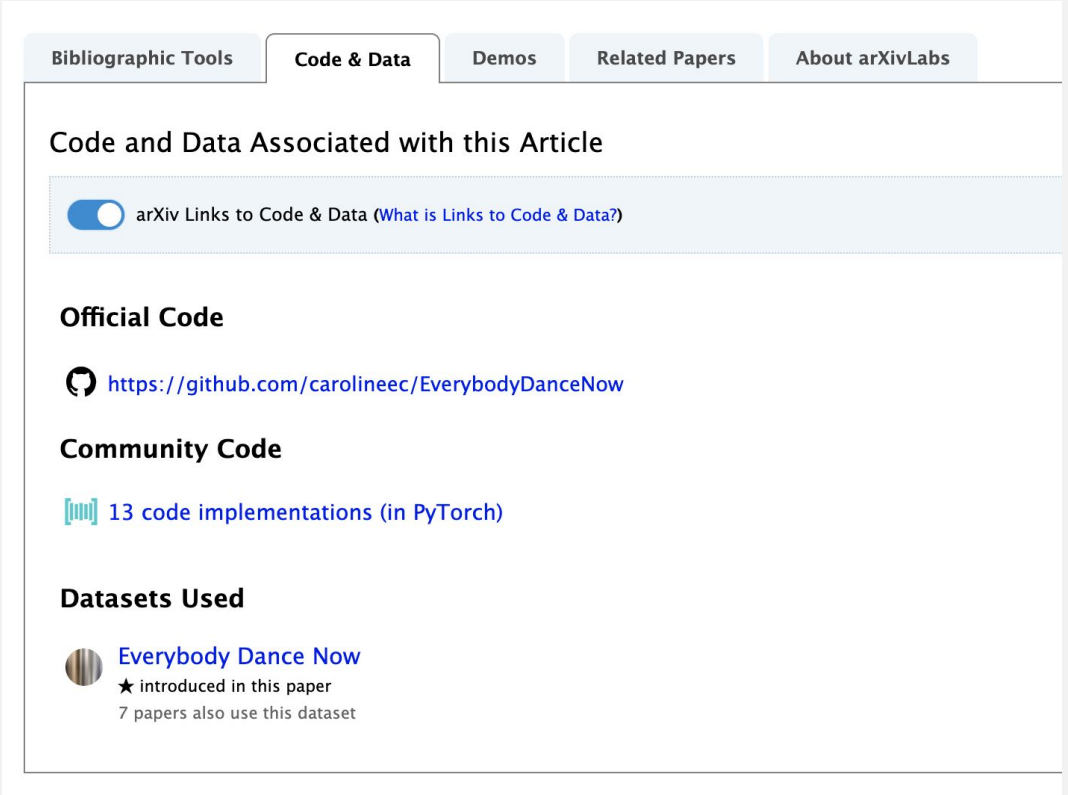
- Largest database of results from published papers



PAPERS WITH CODE

Integrated with:

- arXiv
- ACL anthology
- OpenReview



PAPERS WITH CODE

- Reproducibility reports shown next to original papers

Deep Fair Clustering for Visual Learning

CVPR 2020 · Peizhao Li, Han Zhao, Hongfu Liu · Edit social preview

Fair clustering aims to hide sensitive attributes during data partition by balancing the distribution of protected subgroups in each cluster. Existing work attempts to address this problem by reducing it to a classical balanced clustering with a constraint on the proportion of protected subgroups of the input space...

PDF Abstract

Reproducibility Reports

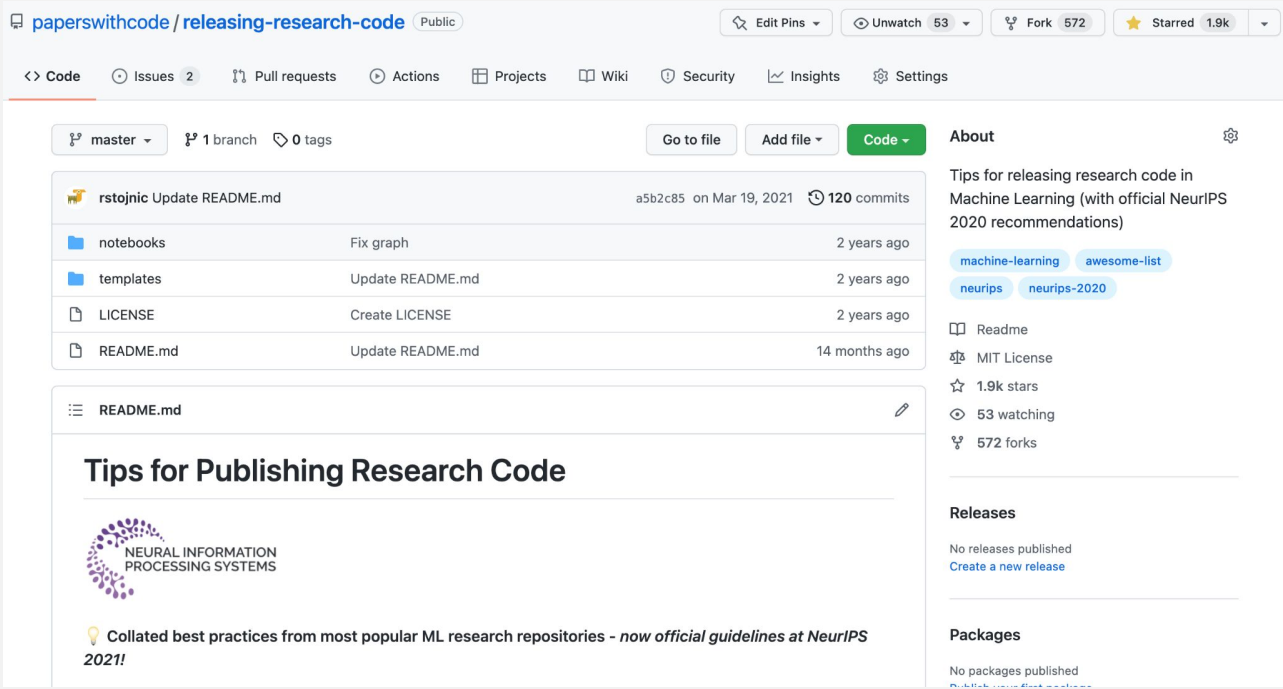
Jan 31 2021 [Re] Deep Fair Clustering for Visual Learning

RC 2020 · Pauline Baanders, Chris Al Gerges, Nienke Reints, Tobias Teule

For the MNIST-USPS dataset, we report similar accuracy and NMI values that are within 1.2% and 0.5% of the values reported in the original paper. However, the balance and entropy differed significantly, where our results were within 73.1% and 30.3% of the original values respectively. For the Color Reverse MNIST dataset, we report similar values on accuracy, balance and entropy, which are within 5.3%, 2.6% and 0.2% respectively. Only the value of the NMI differed significantly, name within 12.9% of the original value In general, our results still support the main claim of the original paper, even though on some metrics the results differ significantly.

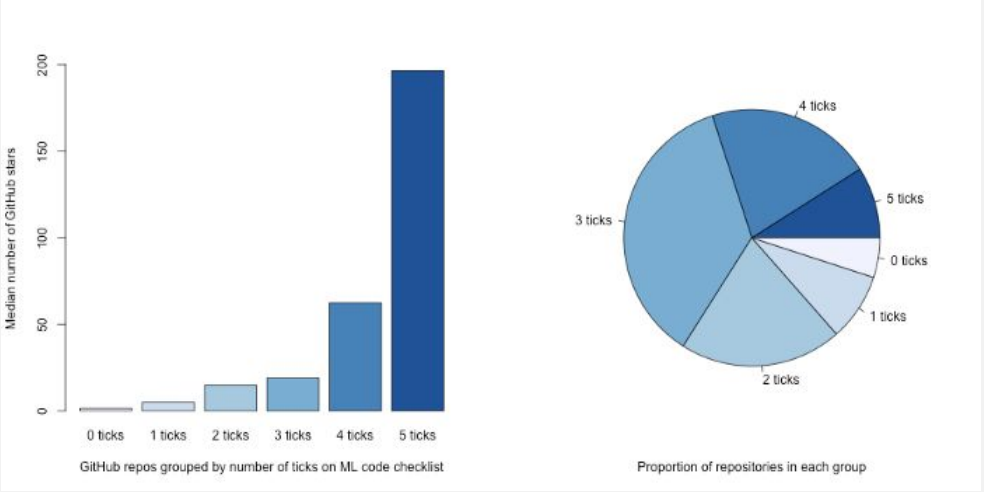
PAPERS WITH CODE

- Collated resources for publishing research code



PAPERS WITH CODE

- ML Code Completeness Checklist (Robert Stojnic, 2020)

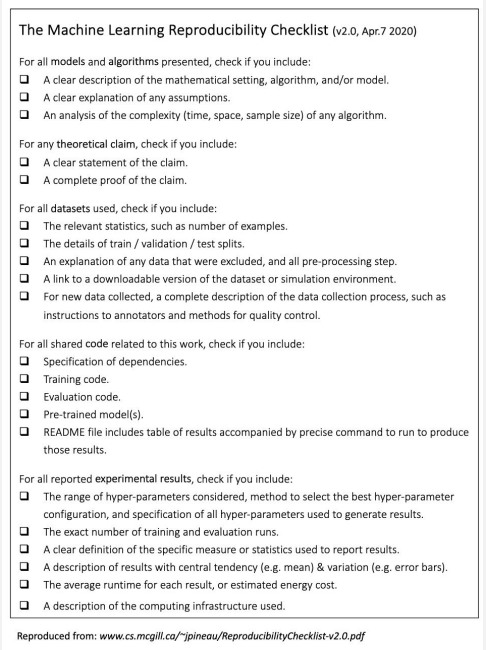


1. **Dependencies** — does a repository have information on dependencies or instructions on how to set up the environment?
2. **Training scripts** — does a repository contain a way to train/fit the model(s) described in the paper?
3. **Evaluation scripts** — does a repository contain a script to calculate the performance of the trained model(s) or run experiments on models?
4. **Pretrained models** — does a repository provide free access to pretrained model weights?
5. **Results** — does a repository contain a table/plot of main results and a script to reproduce those results?

QUESTIONS?

REPRODUCIBILITY CHECKLISTS

- ML Reproducibility Checklist (Joelle Pineau, 2018)
- Minimal information that should be in a manuscript
- Not necessarily exhaustive
- Part of guidelines for major conferences (NeurIPS, ICML, ICLR)



The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020)

For all **models** and **algorithms** presented, check if you include:
- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ A clear explanation of any assumptions.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.

For any **theoretical claim**, check if you include:
- ❑ A clear statement of the claim.
- ❑ A complete proof of the claim.

For all **datasets** used, check if you include:
- ❑ The relevant statistics, such as number of examples.
- ❑ The details of train / validation / test splits.
- ❑ An explanation of any data that were excluded, and all pre-processing step.
- ❑ A link to a downloadable version of the dataset or simulation environment.
- ❑ For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.

For all shared **code** related to this work, check if you include:
- ❑ Specification of dependencies.
- ❑ Training code.
- ❑ Evaluation code.
- ❑ Pre-trained model(s).
- ❑ README file includes table of results accompanied by precise command to run to produce those results.

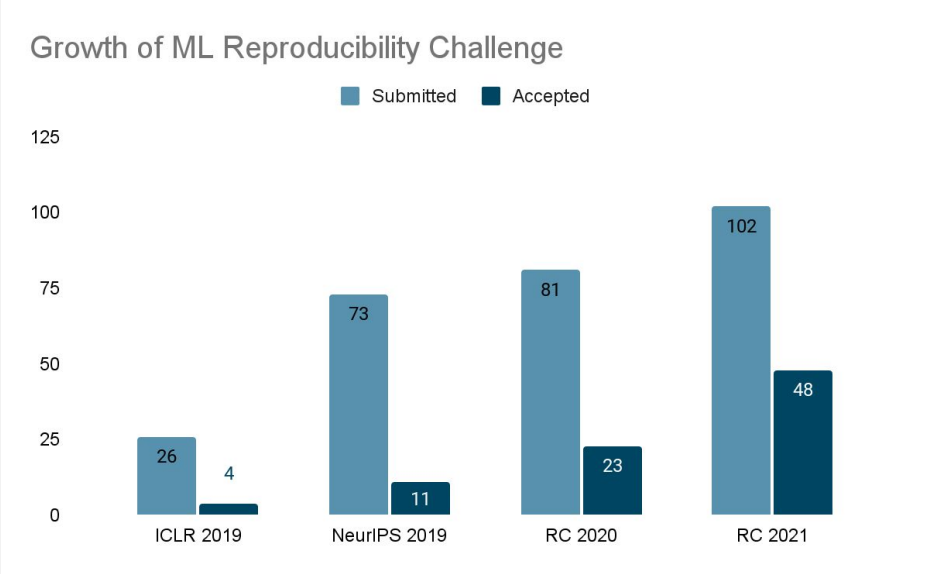
For all reported **experimental results**, check if you include:
- ❑ The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- ❑ The exact number of training and evaluation runs.
- ❑ A clear definition of the specific measure or statistics used to report results.
- ❑ A description of results with central tendency (e.g. mean) & variation (e.g. error bars).
- ❑ The average runtime for each result, or estimated energy cost.
- ❑ A description of the computing infrastructure used.

Reproduced from: *www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf*

REPRODUCIBILITY CHALLENGE

- Started 2018, till date five editions: ICLR 2018, ICLR 2019, NeurIPS 2019, RC 2020, RC 2021
- Task: Choose a submitted paper from a conference, reproduce the central claim of the paper

REPRODUCIBILITY CHALLENGE



REPRODUCIBILITY CHALLENGE

Best Paper Award

▶ **Reproducibility Study of "Counterfactual Generative Networks"**, *Piyush Bagad, Jesse Maas, Paul Hilders, Danilo de Goede*, Forum, Original Paper (ICML 2021)

Outstanding Paper Awards

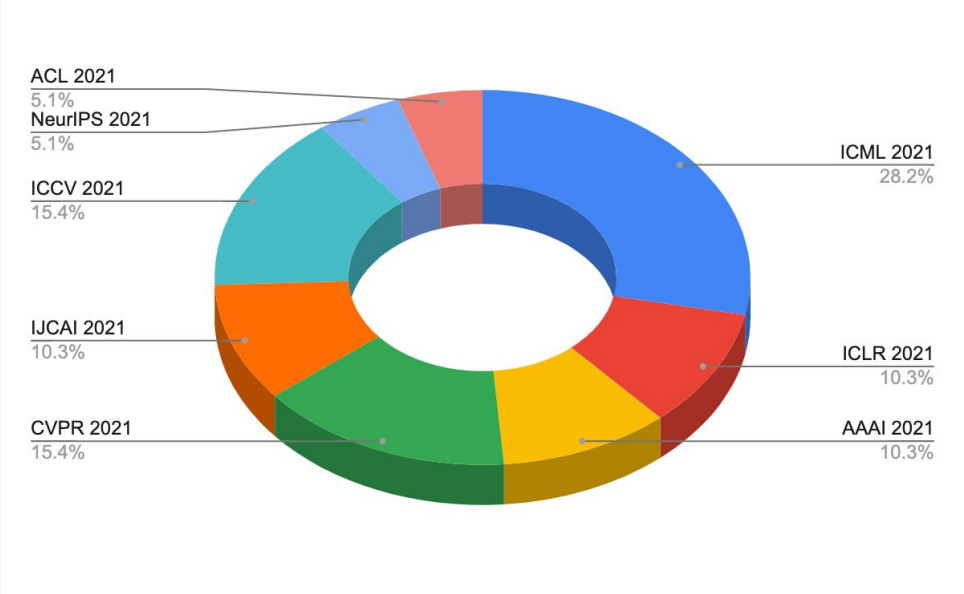
▶ **[Re] Learning to count everything**, *Matija Teršek, Domen Vreš, Maša Kljun*, Forum, Original Paper (CVPR 2021)

▶ **[RE] An Implementation of Fair Robust Learning** , *Ian Hardy*, Forum, Original Paper (ICML 2021)

▶ **Strategic classification made practical: reproduction**, *Guilly Kolkman, Maks kulicki, Jan Athmer, Alex Labro*, Forum, Original Paper (ICML 2021)

▶ **On the reproducibility of "Exacerbating Algorithmic Bias through Fairness Attacks"**, *Andrea Lombardo, Matteo Tafuro, Tin Hadži Veljković, Lasse Becker-Czarnetzki*, Forum, Original Paper (AAAI 2021)

REPRODUCIBILITY CHALLENGE



Reproducibility Reports accepted to MLRC 2021 by conference

REPRODUCIBILITY CHALLENGE

Volume 7 (2021)

Issue 2 (ML Reproducibility Challenge 2020)

1. **Replication in ML Reproducibility Challenge 2020 (Python)** | 10.5281/zenodo.4835602 | PDF | Code | Review | BibTeX
   Verma, R., Wagemans, J.J.O., Dahal, P., and Elfrink, A. 2021. [Re] Explaining Groups of Points in Low-Dimensional Representations. *ReScience C 7*, 2, #24.

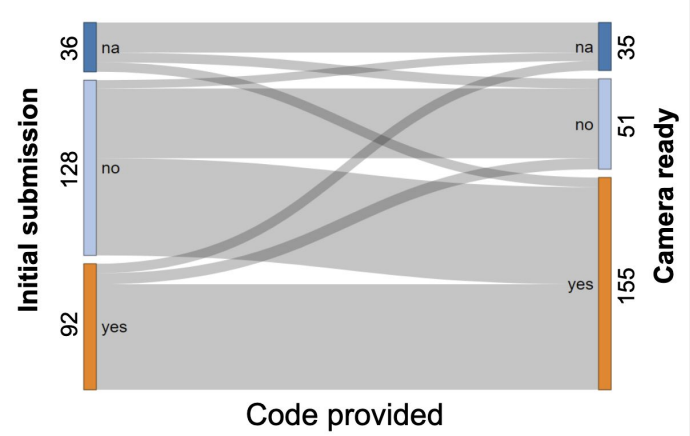
2. **Replication in ML Reproducibility Challenge 2020 (Python)** | 10.5281/zenodo.4833219 | PDF | Code | Data | Review | BibTeX
   Albanis, G., Zioulis, N., Chatzitofis, A., Dimou, A., Zarpalas, D., and Daras, P. 2021. [Re] On end-to-end 6DoF object pose estimation and robustness to object scale. *ReScience C 7*, 2, #2.

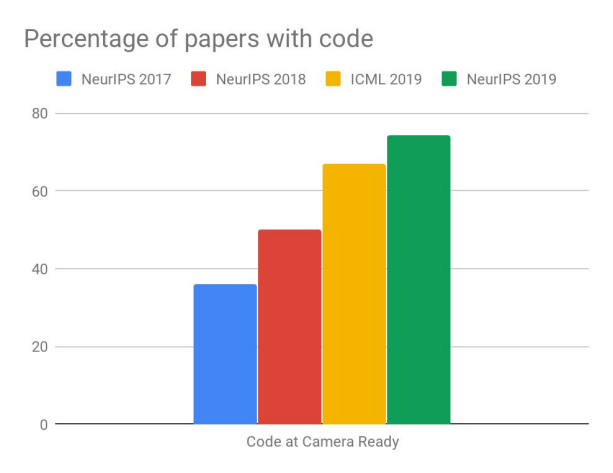
3. **Replication in ML Reproducibility Challenge 2020 (python)** | 10.5281/zenodo.4833389 | PDF | Code | Review | BibTeX
   Arvind, M. and Mama, M. 2021. [Re] Neural Networks Fail to Learn Periodic Functions and How to Fix It. *ReScience C 7*, 2, #3.

RESCIENCE C

IMPACT OF CHECKLISTS AND CHALLENGES

- Increase in the amount of code released during submission
- Increased interaction with authors and practitioners after paper publication through OpenReview





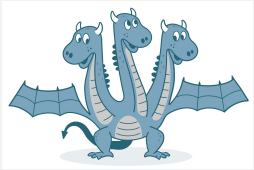
USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release

Link to our previous blog post: https://bit.ly/3LoSuKC

USEFUL TOOLS AND LIBRARIES

- **Config management**
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



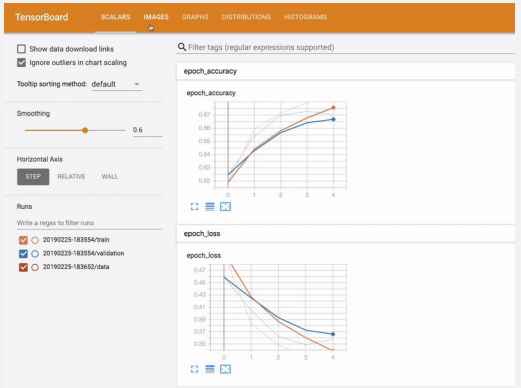
Hydra: https://hydra.cc

Or even plain YAML / JSON files work!

```yaml
general:
  batch_size: 128
  data_name: fashionmnist
  description: This is a sample config
  device: cuda
  epochs: 20
  resume: false
logbook:
  dir: /path/to/log
  logger_file_path: log.jsonl
  log_interval: 100
  project_name: fancy_project
model:
  class_order: 0,1,2,3,4,5,6,7,8,9
  loss_policy: recon_bce # ce, recon_ce, recon_mse, bce, recon_bce
  max_class: 10
  reset_optim: False
  optim:
    eps: 1.0e-08
    learning_rate: 0.001
    name: Adam
    scheduler_gamma: 0.999
    scheduler_patience: 10
    scheduler_type: plateau
    weight_decay: 0.0
  sample_mode: max
  vae_hidden_dim: 50
  z_dim: 5
  resnet:
    in_channels: 1
```

USEFUL TOOLS AND LIBRARIES

- Experimental Config management
- **Logging**
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release

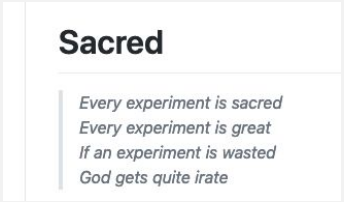


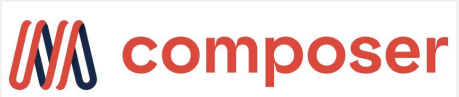
Tensorboard

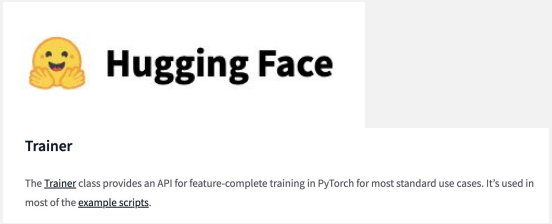
Weights & Biases

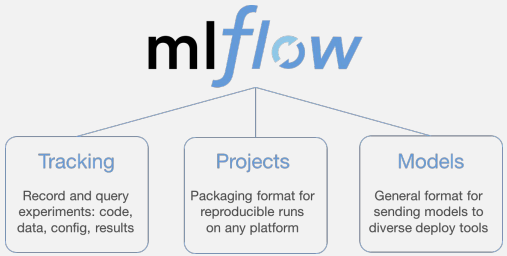
USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- **Experimental Management**
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



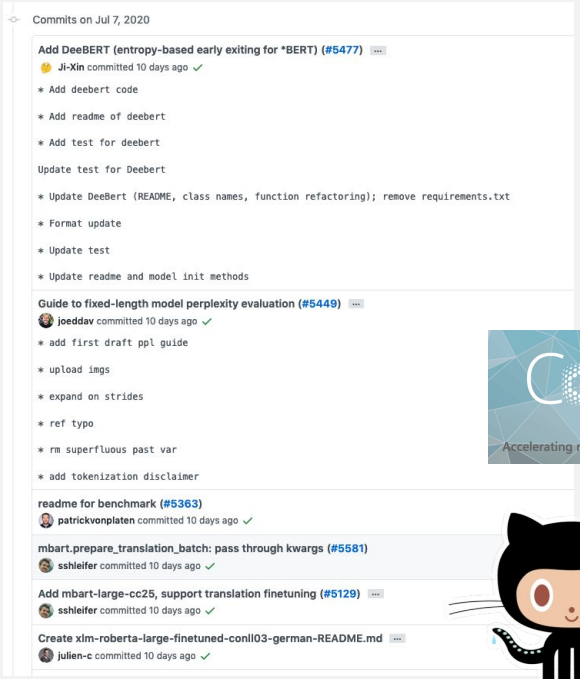






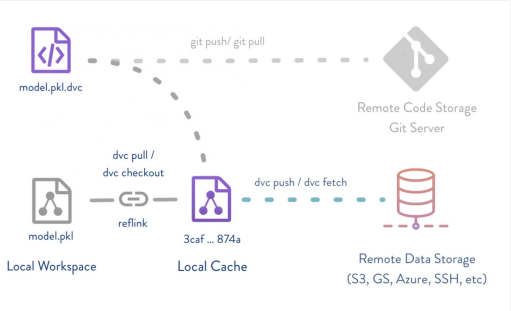
USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- **Versioning**
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- **Data management**
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



DVC, https://dvc.org/

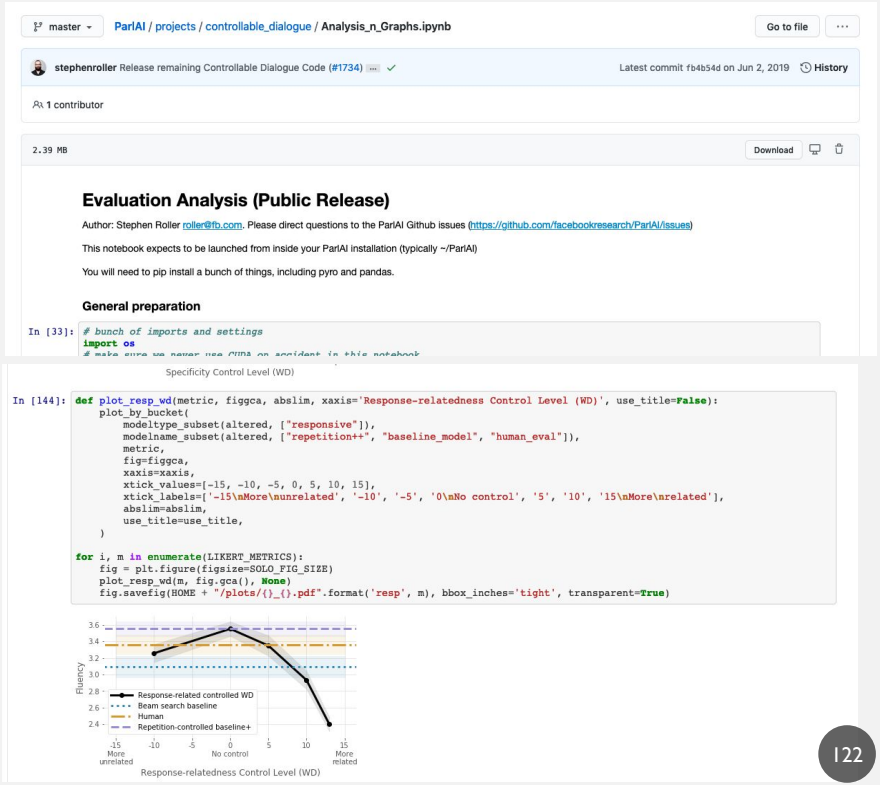
Datasets

Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

USEFUL TOOLS AND LIBRARIES

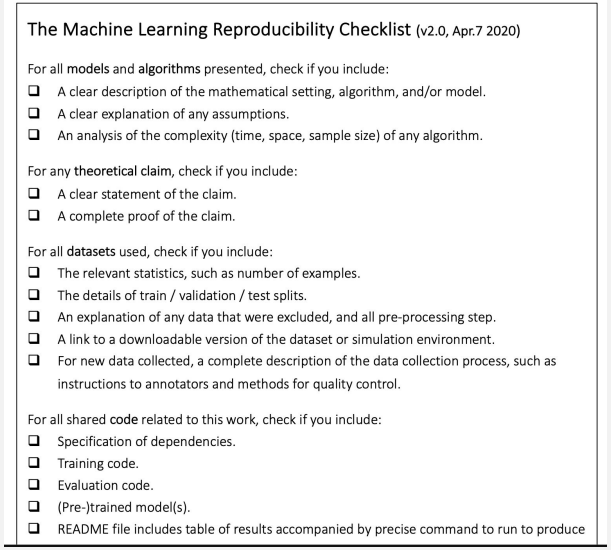
- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- **Data analysis**
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



Relevant works:
https://github.com/EleutherAI/lm-evaluation-harness

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- **Reporting**
- Dependency Management
- Open Source Release
- Effective Communication
- Test and Release



The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020)

For all **models** and **algorithms** presented, check if you include:

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ A clear explanation of any assumptions.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.

For any **theoretical claim**, check if you include:

- ❑ A clear statement of the claim.
- ❑ A complete proof of the claim.

For all **datasets** used, check if you include:

- ❑ The relevant statistics, such as number of examples.
- ❑ The details of train / validation / test splits.
- ❑ An explanation of any data that were excluded, and all pre-processing step.
- ❑ A link to a downloadable version of the dataset or simulation environment.
- ❑ For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.

For all shared **code** related to this work, check if you include:

- ❑ Specification of dependencies.
- ❑ Training code.
- ❑ Evaluation code.
- ❑ (Pre-)trained model(s).
- ❑ README file includes table of results accompanied by precise command to run to produce

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru

{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com

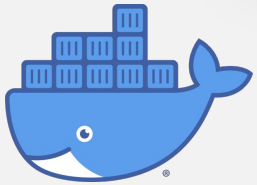
deborah.raji@mail.utoronto.ca

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- **Dependency Management**
- Open Source Release
- Effective Communication
- Test and Release

ReproZip

```
$ pip freeze > requirements.txt
```

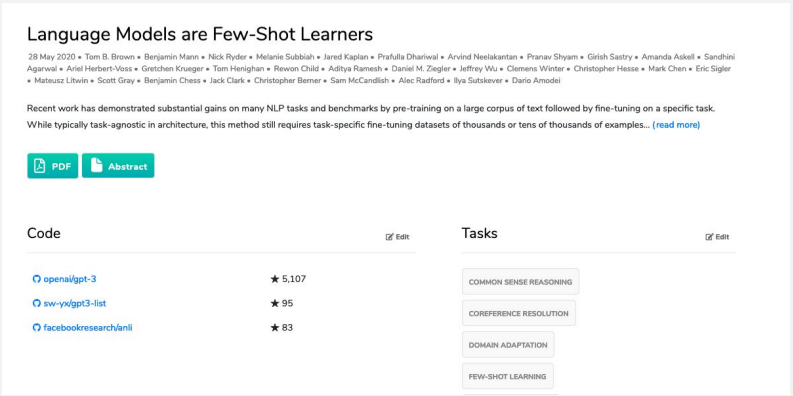


repo2docker

Forde et al,. Reproducible research environments with repo2docker. 2018.

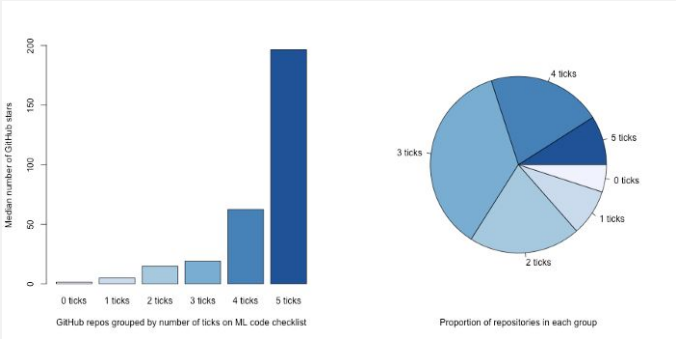
USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- **Open Source Release**
- Effective Communication
- Test and Release



USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- **Effective Communication**
- Test and Release



NeurIPS 2019 repositories with 0 ticks had a median of 1.5 GitHub stars. In contrast, repositories with 5 ticks had a median of 196.5 GitHub stars. Only 9% of repositories had 5 ticks, and most repositories (70%) had 3 ticks or less.

USEFUL TOOLS AND LIBRARIES

- Config management
- Logging
- Experimental Management
- Versioning
- Data management
- Data analysis
- Reporting
- Dependency Management
- Open Source Release
- Effective Communication
- **Test and Release**

Google Colab

QUESTIONS?

MOTIVATION

How can we mitigate the challenges of bigger, more complex models without reducing the benefits?

In this tutorial, we focus on the challenge of ensuring research results are reproducible

TUTORIAL OVERVIEW

1. Introduction to Reproducibility
2. Reproducibility in NLP
3. Mechanisms for Reproducibility
4. Reproducibility as a Teaching Tool