arXiv metadata snapshot

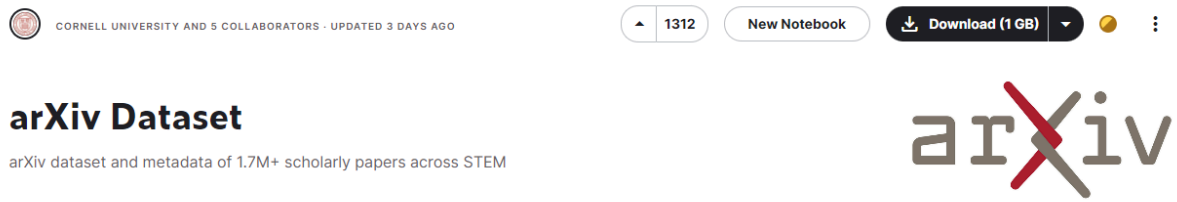
0. 작성 기록

- 2024-05-29
    - 최초 작성

1. aiXiv 논문의 metadata만 저장된 json 파일 다운로드

다운로드 링크

- https://www.kaggle.com/datasets/Cornell-University/arxiv



"**Download (1GB)**" 버튼으로 다운로드 가능

파일 내용 (주요 metadata key 값)

- id : 아카이브 id → pdf 주소 생성에 사용 가능
- authors : 작성자
- title : 제목
- abstract : 요약문 → db 구성, 키워드 생성에 활용
- update_date : 최종 갱신일

파일 내용 (일부)

- 하나의 metadata는 한 줄로 작성됨
- 한 줄의 문장의 시작은 "**{**", 끝은 "**}**"으로 작성됨 (**JSON** 양식)

```
2025172  {"id":"2403.08053","submitter":"Anmol Singhal","authors":"Anmol Singhal, Chirag Jain, Preethu
2025173  {"id":"2403.08054","submitter":"Tzu-Yuan Huang","authors":"Tzu-Yuan Huang, Xiaobing Dai, Sihua
2025174  {"id":"2403.08055","submitter":"Mohamed Elrefaie","authors":"Mohamed Elrefaie, Angela Dai, Fae
2025175  {"id":"2403.08056","submitter":"Martin Berger","authors":"Luke Panayi, Rohan Gandhi, Jim Whitt
2025176  {"id":"2403.08057","submitter":"Hyunsung Cho","authors":"Hyunsung Cho, Yukang Yan, Kashyap Tod
```

… (생략)
{"id":"2403.08054","submitter":"Tzu-Yuan Huang","authors":"Tzu-Yuan Huang, Xiaobing Dai, Sihua Zhang, Alexandre Capone, Velimir\n Todorovski, Stefan Sosnowski and Sandra Hirche","title":"Learning-based Prescribed-Time Safety for Control of Unknown Systems\n with Control Barrier Functions","comments":null,"journal-ref":null,"doi":null,"report-no":null,"categories":"eess.SY cs.SY","license":"http://arxiv.org/licenses/nonexclusive-distrib/1.0/","abstract":" In many control system applications, state constraint satisfaction needs to\nbe guaranteed within a prescribed time. While this issue has been partially\naddressed for systems with known dynamics, it remains largely unaddressed for\nsystems with unknown dynamics. In this paper, we propose a Gaussian\nprocess-based time-varying control method that leverages backstepping and\ncontrol barrier functions

to achieve safety requirements within prescribed time\nwindows. It can be used to keep a system within a safe region or to make it\nreturn to a safe region within a limited time window. These properties are\ncemented by rigorous theoretical results. The effectiveness of the proposed\ncontroller is demonstrated in a simulation of a robotic manipulator.\n","versions":[{"version":"v1","created":"Tue, 12 Mar 2024 20:01:36 GMT"}],"update_date":"2024-03-14","authors_parsed":[["Huang","Tzu-Yuan",""],["Dai","Xiaobing",""],["Zhang","Sihua",""],["Capone","Alexandre",""],["Todorovski","Velimir",""],["Sosnowski","Stefan",""],["Hirche","Sandra",""]]}

{"id":"2403.08055","submitter":"Mohamed Elrefaie","authors":"Mohamed Elrefaie, Angela Dai, Faez Ahmed","title":"DrivAerNet: A Parametric Car Dataset for Data-Driven Aerodynamic Design\n and Graph-Based Drag Prediction","comments":null,"journal-ref":null,"doi":null,"report-no":null,"categories":"cs.LG physics.flu-dyn","license":"http://creativecommons.org/licenses/by/4.0/","abstract":" This study introduces DrivAerNet, a large-scale high-fidelity CFD dataset of\n3D industry-standard car shapes, and RegDGCNN, a dynamic graph convolutional\nneural network model, both aimed at aerodynamic car design through machine\nlearning. DrivAerNet, with its 4000 detailed 3D car meshes using 0.5 million\nsurface mesh faces and comprehensive aerodynamic performance data comprising of\nfull 3D pressure, velocity fields, and wall-shear stresses, addresses the\ncritical need for extensive datasets to train deep learning models in\nengineering applications. It is 60\\% larger than the previously available\nlargest public dataset of cars, and is the only open-source dataset that also\nmodels wheels and underbody. RegDGCNN leverages this large-scale dataset to\nprovide high-precision drag estimates directly from 3D meshes, bypassing\ntraditional limitations such as the need for 2D image rendering or Signed\nDistance Fields (SDF). By enabling fast drag estimation in seconds, RegDGCNN\nfacilitates rapid aerodynamic assessments, offering a substantial leap towards\nintegrating data-driven methods in automotive design. Together, DrivAerNet and\nRegDGCNN promise to accelerate the car design process and contribute to the\ndevelopment of more efficient vehicles. To lay the groundwork for future\ninnovations in the field, the dataset and code used in our study are publicly\naccessible at \\url{https://github.com/MohamedElrefaie/DrivAerNet}\n","versions":[{"version":"v1","created":"Tue, 12 Mar 2024 20:02:39 GMT"}],"update_date":"2024-03-14","authors_parsed":[["Elrefaie","Mohamed",""],["Dai","Angela",""],["Ahmed","Faez",""]]}

{"id":"2403.08056","submitter":"Martin Berger","authors":"Luke Panayi, Rohan Gandhi, Jim Whittaker, Vassilios Chouliaras, Martin\n Berger, Paul Kelly","title":"Improving Memory Dependence Prediction with Static Analysis","comments":"15 pages","journal-ref":null,"doi":null,"report-no":null,"categories":"cs.PL cs.AR","license":"http://creativecommons.org/licenses/by-nc-nd/4.0/","abstract":" This paper explores the potential of communicating information gained by\nstatic analysis from compilers to Out-of-Order (OoO) machines, focusing on the\nmemory dependence predictor (MDP). The MDP enables loads to issue without all\nin-flight store addresses being known, with minimal memory order violations. We\nuse LLVM to find loads with no dependencies and label them via their opcode.\nThese labelled loads skip making lookups into the MDP, improving prediction\naccuracy by reducing false dependencies. We communicate this information in a\nminimally intrusive way, i.e.~without introducing additional hardware costs or\ninstruction bandwidth, providing these improvements without any additional\noverhead in the CPU. We find that in select cases in Spec2017, a significant\nnumber of load instructions can skip interacting with the MDP and lead to a\nperformance gain. These results point to greater possibilities for static\nanalysis as a source of near zero cost performance gains in future CPU designs.\n","versions":[{"version":"v1","created":"Tue, 12 Mar 2024 20:04:09 GMT"},{"version":"v2","created":"Sat, 4 May 2024 20:03:32 GMT"}],"update_date":"2024-05-07","authors_parsed":[["Panayi","Luke",""],["Gandhi","Rohan",""],["Whittaker","Jim",""],["Chouliaras","Vassilios",""],["Berger","Martin",""],["Kelly","Paul",""]]}

… (생략)

2. aiXiv 논문 pdf 파일 다운로드 (bulk)

24년 03월 논문의 PDF 전체를 다운로드하는 경우, 아래의 명령어 사용 가능
※ colab 활용 가능

$	gsutil cp -r gs://arxiv-dataset/arxiv/arxiv/pdf/2403/ ./papers_from_2024/

※ 로컬에 “papers_from_2024” 폴더가 존재하지 않은 경우, 다운로드 불가.
폴더는 미리 만들것