PublicDatasets-Ahmet

Files & media	
Github	https://github.com/madprogramer/PublicDatasets
Other authors	
Overleaf	https://www.overleaf.com/3623354424hwmfyndjgvbc
Report	
Status	Doing
Type	MSc-7.5
When	Fall 2022
Who	Ahmet

PublicDatasets

How do datasets sneak into journal papers?

At a glance

Machine learning challenges hosted on platforms such as Kaggle, grand-challenge.org, or CodaLab have attracted a lot of attention, both from academia and industry researchers. Challenge designs vary widely, including what type of data is available, how the algorithms are evaluated, and the rewards for the winners. In medical imaging, there is some evidence that challenges might be creating a shift in attention to different diseases, for example, there is a disproportionate increase in papers on lung cancer after the 2016 Kaggle challenge on the subject.

However, it is actually non-trivial to observe the impact these datasets have on research and technology. There are no well-defined conventions for citing most public datasets, which often come from outside of academic journals. The overall goal of this project is to understand how we can relate research to datasets they either use or are inspired by.

Log

Upcoming:

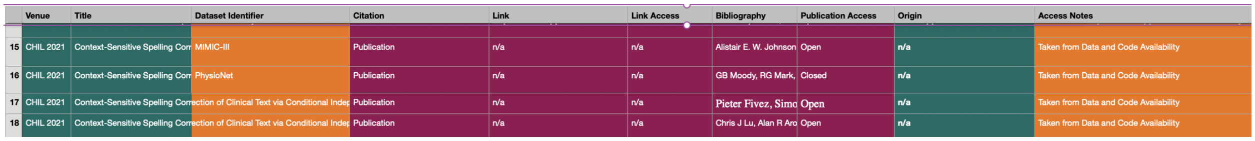
- Ask Veronika if it’s a good idea to include RAW PDFs in dataset, or to only link to proceedings?
- Manual Dataset column, Manual Evidence, Other notes.
    - You can use identifiers or categories
    - Try to associate with venue
- Identifier Extraction from Keywords
- Add multiple rows for papers mentioning several datasets.
- Invivo, taguette
- Imagine what your plots should look like
    - Barchart of `kaggle`, `not given` etc.
    - https://github.com/christinekaarde/Value-Analysis-Thesis/blob/main/04-analysis/data-analysis-(quantitative).ipynb

28 November 2022

- Join the figure (2 images in 1 figure**)
- Enlarge the fontsize

18 November 2022: How the turn tables

- Updated fields.



	Venue	Title	Dataset Identifier	Citation	Link	Link Access	Bibliography	Publication Access	Origin	Access Notes
15	CHIL 2021	Context-Sensitive Spelling Corr	MIMIC-III	Publication	n/a	n/a	Alistair E. W. Johnson	Open	n/a	Taken from Data and Code Availability
16	CHIL 2021	Context-Sensitive Spelling Corr	PhysioNet	Publication	n/a	n/a	GB Moody, RG Mark,	Closed	n/a	Taken from Data and Code Availability
17	CHIL 2021	Context-Sensitive Spelling Correction of Clinical Text via Conditional Indep		Publication	n/a	n/a	Pieter Fivez, Simo	Open	n/a	Taken from Data and Code Availability
18	CHIL 2021	Context-Sensitive Spelling Correction of Clinical Text via Conditional Indep		Publication	n/a	n/a	Chris J Lu, Alan R Aro	Open	n/a	Taken from Data and Code Availability

- Will now focus on CHIL because of dedicated `Data and Code` Availability section.
- Report notes:
  - **Indicate possibility of missing data**
  - Absence of Evidence is not Evidence of Absence

Todo for weekend:

- Start Overleaf
- Write first few sections
- Expand dataset for CHIL 2021
- Plot some preliminary results!

17 November 2022: Logs II

Who did you help this week?

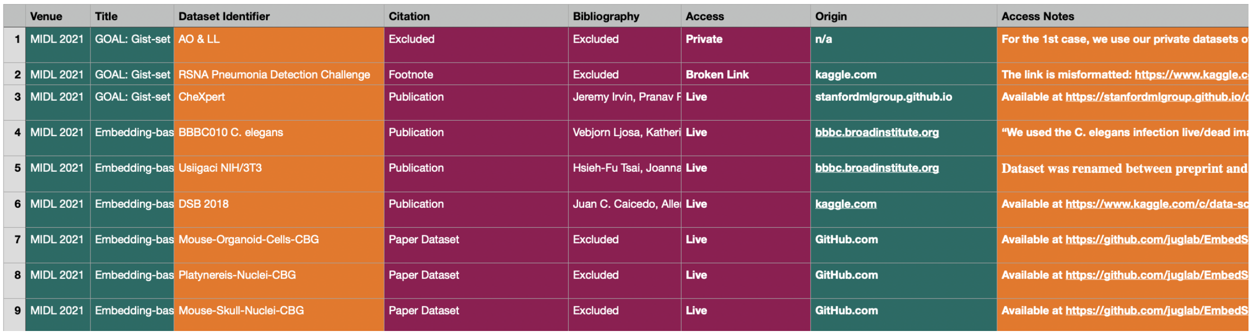
- No one :|

What helped you this week?

- Christina's Thesis Notes have been helpful figuring out what to do next.

What did you achieve?

- Data Schema for Datasets in Research Papers:



	Venue	Title	Dataset Identifier	Citation	Bibliography	Access	Origin	Access Notes
1	MIDL 2021	GOAL: Gist-set	AO & LL	Excluded	Excluded	Private	n/a	For the 1st case, we use our private datasets o
2	MIDL 2021	GOAL: Gist-set	RSNA Pneumonia Detection Challenge	Footnote	Excluded	Broken Link	kaggle.com	The link is misformatted: https://www.kaggle.c
3	MIDL 2021	GOAL: Gist-set	CheXpert	Publication	Jeremy Irvin, Pranav F	Live	stanfordmlgroup.github.io	Available at https://stanfordmlgroup.github.io/c
4	MIDL 2021	Embedding-bas	BBBC010 C. elegans	Publication	Vebjorn Ljosa, Katheri	Live	bbbc.broadinstitute.org	"We used the C. elegans infection live/dead ima
5	MIDL 2021	Embedding-bas	Usiigaci NIH/3T3	Publication	Hsieh-Fu Tsai, Joanna	Live	bbbc.broadinstitute.org	Dataset was renamed between preprint and
6	MIDL 2021	Embedding-bas	DSB 2018	Publication	Juan C. Caicedo, Aller	Live	kaggle.com	Available at https://www.kaggle.com/c/data-sc
7	MIDL 2021	Embedding-bas	Mouse-Organoid-Cells-CBG	Paper Dataset	Excluded	Live	GitHub.com	Available at https://github.com/juglab/EmbedS
8	MIDL 2021	Embedding-bas	Platynereis-Nuclei-CBG	Paper Dataset	Excluded	Live	GitHub.com	Available at https://github.com/juglab/EmbedS
9	MIDL 2021	Embedding-bas	Mouse-Skull-Nuclei-CBG	Paper Dataset	Excluded	Live	GitHub.com	Available at https://github.com/juglab/EmbedS

- Can fully automate:
  - Venue
  - Title
  - Dataset Origin (given URL)
- Can partially automate (**on keyword matches. F.x. Kaggle.**)
  - Citation Category (Footnote, Journal Publication or URL)
  - Access (Public or Private)

    - Bibliography Mentions
- Not so easy to automate:
- Dataset Identifier
    - Multiple references, sometimes full name sometimes abbreviated.
    - f.x. Usiigaci is not trivially associable with the title `Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl`.
- Notes
    - When keyword match isn't present, annotation requires further context.

What did you struggle with?

- No matter what I do, some manual annotation is needed.
    - This hurts the scalability of my results.
- Behind on plotting, but I expect to catch up next week.
- I realized it might be a bit overkill to include **every** dataset used in a research paper.
    - If I only include challenge datasets, I will only be able to compare them amongst each other:
        - Footnotes vs. Bibliographic Citations. Broken, Rotten and Active Links.
    - But if I should exclude any datasets, I don't know which ones to exclude.
        - Can compare dataset hosts (GitHub vs. Institutional), frequency of challenge datasets against baseline.

What would you like to work on next week?

- Results, results, results!
- If I complete the pipeline on the automatable data,
    - then I can manually fill in the remaining fields and update figures in future weeks.
    - This would require me to churn most of the report within the next 2 weeks.

Where do you need help from Veronika?

- What should my inclusion/exclusion criteria be on datasets from our collected research papers?
- My deadline is December 15, but I am starting to worry if I might need an extension.
- What are my options? Scaling down my scope or extending my deadline?

9 November 2022: Logs I

Who did you help this week?

- No one :|

What helped you this week?

- Reading past notes.

What did you achieve?

- Some notes on challenge dataset hunting:
    - In papers comparing multiple datasets there is often a dedicated section listing all datasets.
    - Appendices matter! There may be additional datasets only mentioned after the paper.
- Challenge datasets are presented in one of three ways:
- The X competition

    - Ex. *moco-cxr performance on the* `chexpert` *competition task pathologies.*
- The X challenge
    - Ex. *the 2019* `fastmri` *challenge*
    - as opposed to *The goal of the conference is to foster excellent research that addresses the unique* `challenge`*s and opportunities*
- URL identifier
    - Ex. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge or https://promise12.grand-challenge.org

What did you struggle with?

- Analysis Paralysis:
    - I'm not sure what I want to annotate, and with what tool
    - How to scale it to 128 PDFs?
- Comparing non-challenge datasets to challenge datasets
    - Annotation seems infeasible.

What would you like to work on next week?

- Improving biblography checker.
- Retrieving contest identifier by pattern-matching

Where do you need help from Veronika?

- Any other querying ideas?
- Good practices for annotation?
- What should I tag and why?
- My hypothesis is that "Challenge datasets are under-represented in bibliographies (than there actually are)"

Any other topics

- Forgot to add checks on Title :_)

31-10-2022: Monday Meeting:

For next week:

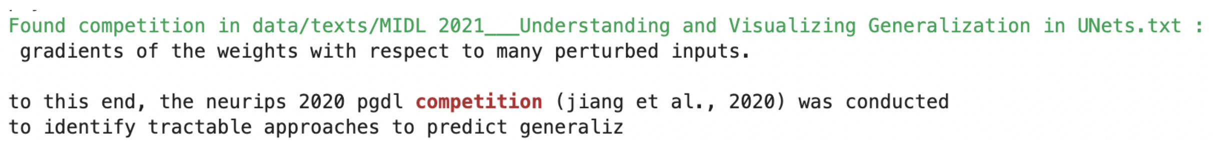
- Report Draft: (will share Overleaf)
    - Venue Sources (Selection Criteria)
    - Table of Rules (Keyword-detection Rules + Recognising Dataset name)
    - Annotate by-hand over subset of reports and compare (n=20, but also **read**)
        - What *extra* information can we benefit from?
            - Domain: Body Part,
            - Modality: Imaging Task,
            - Availability: Private/Public,
            - Origin: Kaggle or Stanford,
            - any in-text explanations?)
    - Other ideas: GitHub links in abstract/body
    - **Important: Focus in-body and in-abstract mentions**

        - External tools can help with parsing bibliography
    - **Important:** Formulate a good hypothesis
        - Try to assure that your results are reliable

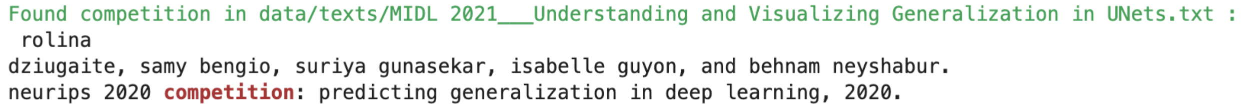
30-10-2022: Organizing

- Now showing first occurrence before and in Bibliography

Example:



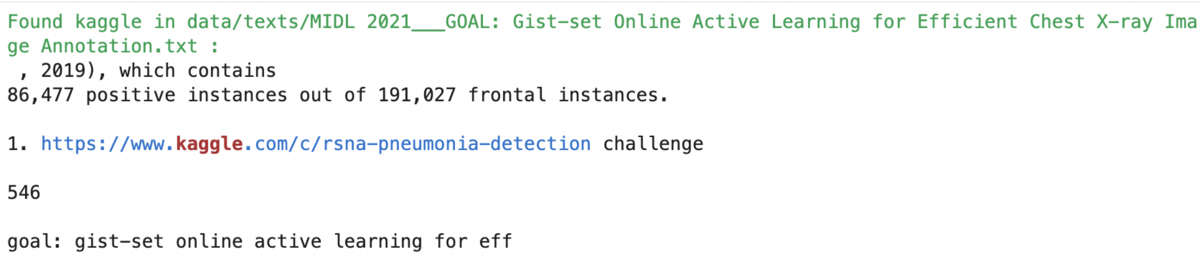
```
Found competition in data/texts/MIDL 2021___Understanding and Visualizing Generalization in UNets.txt :
 gradients of the weights with respect to many perturbed inputs.

to this end, the neurips 2020 pgdl competition (jiang et al., 2020) was conducted
to identify tractable approaches to predict generaliz
```



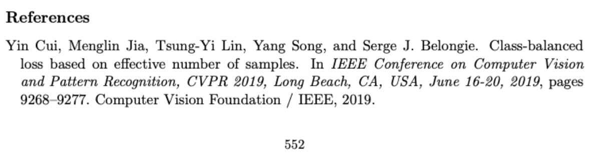
```
Found competition in data/texts/MIDL 2021___Understanding and Visualizing Generalization in UNets.txt :
 rolina
dziugaite, samy bengio, suriya gunasekar, isabelle guyon, and behnam neyshabur.
neurips 2020 competition: predicting generalization in deep learning, 2020.
```

Many competition datasets are footnoted only, with no match in bibliography:

- RSNA Pneumonia Detection challenge in `Disability prediction in multiple sclerosis using performance outcome measures and demographic data.` (CHIL21)



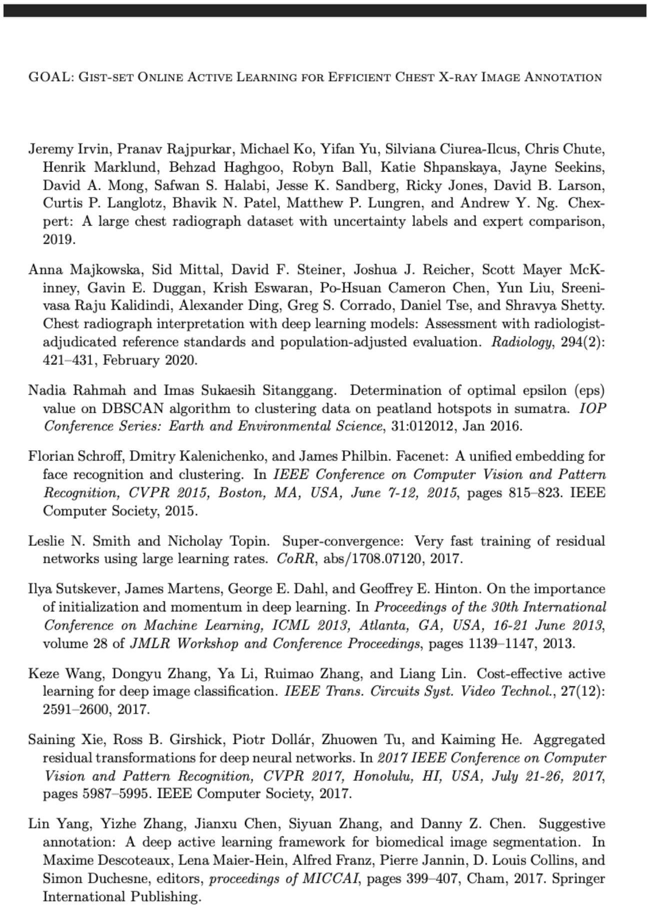
```
Found kaggle in data/texts/MIDL 2021___GOAL: Gist-set Online Active Learning for Efficient Chest X-ray Ima
ge Annotation.txt :
 , 2019), which contains
86,477 positive instances out of 191,027 frontal instances.

1. https://www.kaggle.com/c/rsna-pneumonia-detection challenge

546

goal: gist-set online active learning for eff
```



References

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9268–9277. Computer Vision Foundation / IEEE, 2019.

552



GOAL: Gist-set Online Active Learning for Efficient Chest X-ray Image Annotation

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

Anna Majkowska, Sid Mittal, David F. Steiner, Joshua J. Reicher, Scott Mayer McKinney, Gavin E. Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, Alexander Ding, Greg S. Corrado, Daniel Tse, and Shravya Shetty. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2): 421–431, February 2020.

Nadia Rahmah and Imas Sukaesih Sitanggang. Determination of optimal epsilon (eps) value on DBSCAN algorithm to clustering data on peatland hotspots in sumatra. *IOP Conference Series: Earth and Environmental Science*, 31:012012, Jan 2016.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*, abs/1708.07120, 2017.

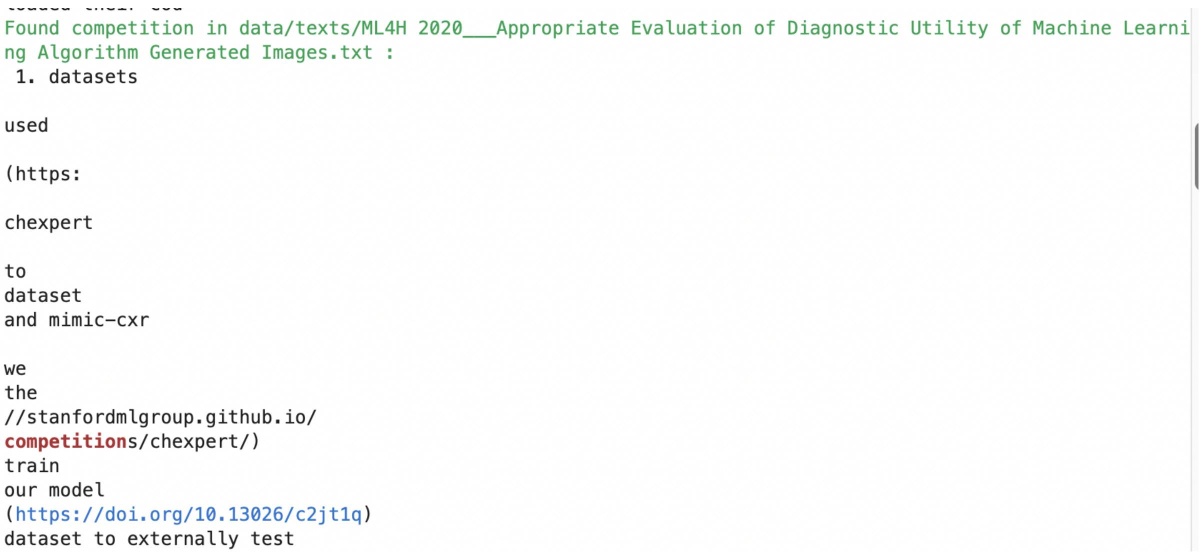
Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1139–1147, 2013.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Trans. Circuits Syst. Video Technol.*, 27(12): 2591–2600, 2017.

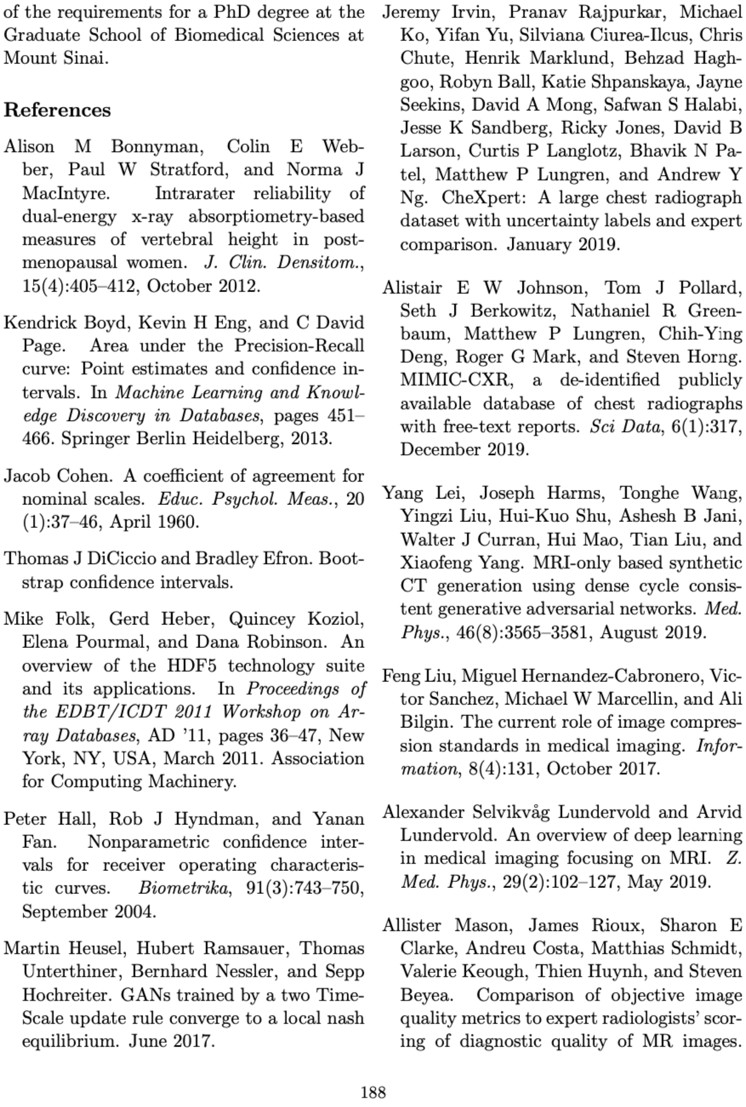
Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995. IEEE Computer Society, 2017.

Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *proceedings of MICCAI*, pages 399–407, Cham, 2017. Springer International Publishing.

- Stanford's chexphoto in `Appropriate Evaluation of Diagnostic Utility of Machine Learning Algorithm Generated Image` (ML4H 2020)



```
Found competition in data/texts/ML4H 2020___Appropriate Evaluation of Diagnostic Utility of Machine Learni
ng Algorithm Generated Images.txt :
 1. datasets

used

(https:

chexpert

to
dataset
and mimic-cxr

we
the
//stanfordmlgroup.github.io/
competitions/chexpert/)
train
our model
(https://doi.org/10.13026/c2jt1q)
dataset to externally test
```



of the requirements for a PhD degree at the Graduate School of Biomedical Sciences at Mount Sinai.

References

Alison M Bonnyman, Colin E Webber, Paul W Stratford, and Norma J MacIntyre. Intrarater reliability of dual-energy x-ray absorptiometry-based measures of vertebral height in postmenopausal women. *J. Clin. Densitom.*, 15(4):405–412, October 2012.

Kendrick Boyd, Kevin H Eng, and C David Page. Area under the Precision-Recall curve: Point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer Berlin Heidelberg, 2013.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20 (1):37–46, April 1960.

Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals.

Mike Folk, Gerd Heber, Quincey Koziol, Elena Pourmal, and Dana Robinson. An overview of the HDF5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*, AD '11, pages 36–47, New York, NY, USA, March 2011. Association for Computing Machinery.

Peter Hall, Rob J Hyndman, and Yanan Fan. Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, 91(3):743–750, September 2004.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two Time-Scale update rule converge to a local nash equilibrium. June 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. January 2019.

Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-Ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, 6(1):317, December 2019.

Yang Lei, Joseph Harms, Tonghe Wang, Yingzi Liu, Hui-Kuo Shu, Ashesh B Jani, Walter J Curran, Hui Mao, Tian Liu, and Xiaofeng Yang. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med. Phys.*, 46(8):3565–3581, August 2019.

Feng Liu, Miguel Hernandez-Cabronero, Victor Sanchez, Michael W Marcellin, and Ali Bilgin. The current role of image compression standards in medical imaging. *Information*, 8(4):131, October 2017.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.*, 29(2):102–127, May 2019.

Allister Mason, James Rioux, Sharon E Clarke, Andreu Costa, Matthias Schmidt, Valerie Keough, Thien Huynh, and Steven Beyea. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images.

29-10-2022: Coding

- Updated notebook to allow for multiple venues (not too difficult)
    - Venues Covered:

28-10-2022: Blogging and Coding

https://medium.com/memex/i-found-it-on-the-internet-the-hidden-impact-of-open-data-af0c1b954602

27-10-2022: Takeaways from Digital Tech Summit Day 2

- Novozymes
- WSAudiology
- Can Data Visualizations Save Lives?
- https://en.wikipedia.org/wiki/Bill_James
- https://www.datarobot.com/wiki/citizen-data-scientist/

26-10-2022: Takeaways from Digital Tech Summit Day 1

- Quantum
- Favorite Talk: James Love (Novozymes automation)
- Other Interesting: https://trifork.com/about/
- Also during the same day I discovered https://www.explainpaper.com/, a very useful forthcoming utility 🙂
- Skilled labor shortage
- Major Problem of Scalability
- Federated Learning is desperately needed.
- Danske Folkepartiet

24-10-2022 : TODOs

- First:
    - Update notebook to allow for multiple venues (not too difficult)
    - Changes to iterator:
        - Include a couple of new venues.
    - Changes to keywords:
        - Try variations on dataset-hunting keywords.
    - Modularise that notebook!
        - Suggestion: https://github.com/adriapr/crowdairway
    - Add a `show_context` function to show snippet of occurrences.
- Further Ideas:
    - **It is rare, but possible**, for a keyword to only appear in the title instead of the body.
        - **Check for both!**
    - Look at the `lung cancer` matches to see if there are any more “hidden” references to `DSB 2017`.
    - Frequency of matches instead of 0/1.
    - Difficult but not impossible:
        - Try to compare URL citations against bibliography!

07-10-2022 : Christine Meeting Notes

- Drop DOIs, you don't need them.
    - Ask Veronika if it's a good idea to include RAW PDFs in dataset, or to only link to proceedings?
- **title scanning for keywords was a pain**
    - Wish we talked sooner to tell about `pdfminer`
- Springer makes things easier if you do MICCAI in the future.
- **Articulated my problem:**
    - **Find *how* Kaggle challenges are being mentioned**
        - **Link-only**
        - **Abbreviation**
        - **Bibliography reference**

06-10-2022 : October Agenda

- Have a separate `papers.csv` which only contains information on the papers and venue (minus keyword matches).
    - This approach gives flexibility for alternative match rules.
        - An additional `.csv` can be generated based on alternative rules.
    - Add a "keyword_" prefix to mark keyword fields.
        - Also "keyword_intitle", "keyword_intext" to inform about locality.
        - **Extend function to show keyword in context.**
    - Talk to Christine about scraping MICCAI and additional venues.
- **It is rare, but possible**, for a keyword to only appear in the title instead of the body.
    - **Check for both!**
- Experiment: "Can we find `grand-challenge`?"
    - We actually found 2 results:

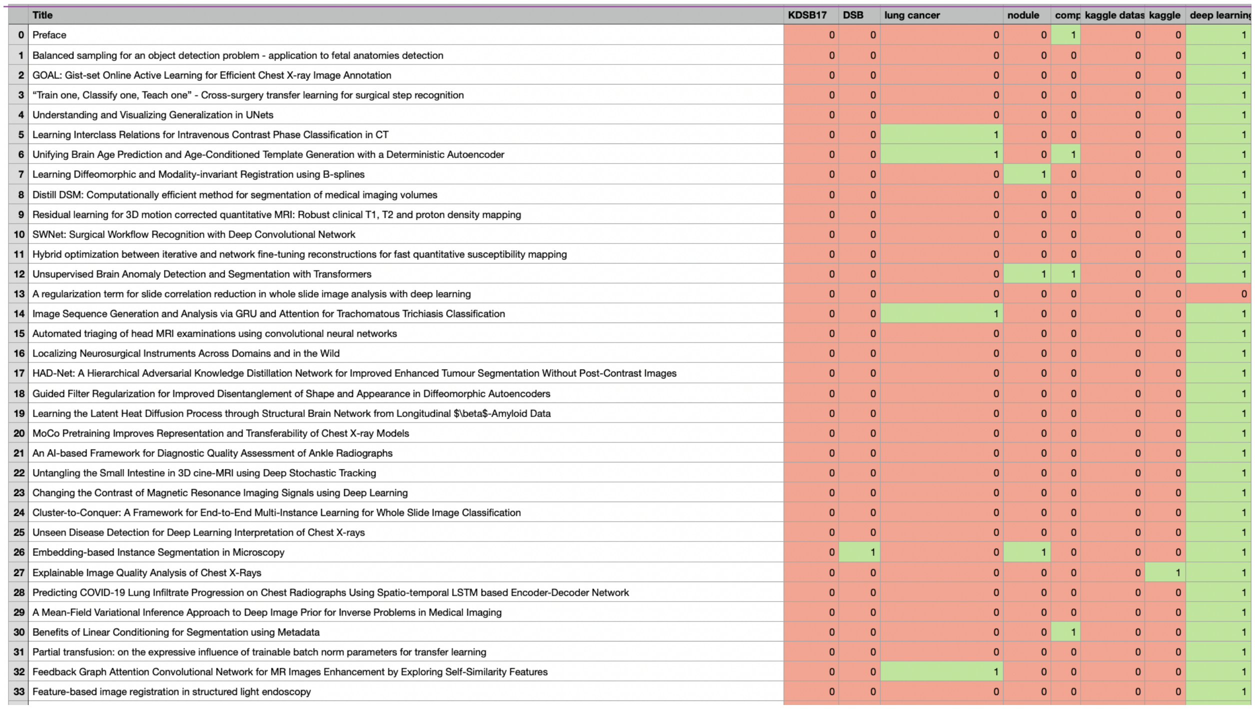
```
Found grand-challenge in Weakly Supervised Volumetric Segmentation via Self-taught Shape Denoising Model : "promise12.grand-chal..."
Found grand-challenge in Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification :
```

- Some other dataset-hunting keywords:
    - Original URL
    - `public * available`
        - Regular expression search also a good extension!
    - `public data`
- Discovery:
    - My crawling pattern works on anything from `proceedings.mlr.press`
    - We can then try some other datasets from ML/CV conferences
        - CHIL 2021
- Takeaways:
    - Update notebook to allow for multiple venues (not too difficult)
    - Changes to iterator:

        - Include a couple of new venues.
    - Changes to keywords:
        - Try variations on dataset-hunting keywords.
    - Modularise that notebook!
        - Suggestion: https://github.com/adriapr/crowdairway
    - Add a `show_context` function to show snippet of occurrences.
- Further Ideas:
    - Look at the `lung cancer` matches to see if there are any more “hidden” references to `DSB 2017`.
    - Frequency of matches instead of 0/1.
    - Difficult but not impossible:
        - Try to compare URL citations against bibliography!

05-10-2022 : Matchmaking

Output at last! Today was a busy day (^_^)



	Title	KDSB17	DSB	lung cancer	nodule	comp	kaggle datas	kaggle	deep learnin
0	Preface	0	0	0	0	1	0	0	1
1	Balanced sampling for an object detection problem - application to fetal anatomies detection	0	0	0	0	0	0	0	1
2	GOAL: Gist-set Online Active Learning for Efficient Chest X-ray Image Annotation	0	0	0	0	0	0	0	1
3	“Train one, Classify one, Teach one” - Cross-surgery transfer learning for surgical step recognition	0	0	0	0	0	0	0	1
4	Understanding and Visualizing Generalization in UNets	0	0	0	0	0	0	0	1
5	Learning Interclass Relations for Intravenous Contrast Phase Classification in CT	0	0	1	0	0	0	0	1
6	Unifying Brain Age Prediction and Age-Conditioned Template Generation with a Deterministic Autoencoder	0	0	1	0	1	0	0	1
7	Learning Diffeomorphic and Modality-invariant Registration using B-splines	0	0	0	1	0	0	0	1
8	Distill DSM: Computationally efficient method for segmentation of medical imaging volumes	0	0	0	0	0	0	0	1
9	Residual learning for 3D motion corrected quantitative MRI: Robust clinical T1, T2 and proton density mapping	0	0	0	0	0	0	0	1
10	SWNet: Surgical Workflow Recognition with Deep Convolutional Network	0	0	0	0	0	0	0	1
11	Hybrid optimization between iterative and network fine-tuning reconstructions for fast quantitative susceptibility mapping	0	0	0	0	0	0	0	1
12	Unsupervised Brain Anomaly Detection and Segmentation with Transformers	0	0	0	1	1	0	0	1
13	A regularization term for slide correlation reduction in whole slide image analysis with deep learning	0	0	0	0	0	0	0	0
14	Image Sequence Generation and Analysis via GRU and Attention for Trachomatous Trichiasis Classification	0	0	1	0	0	0	0	1
15	Automated triaging of head MRI examinations using convolutional neural networks	0	0	0	0	0	0	0	1
16	Localizing Neurosurgical Instruments Across Domains and in the Wild	0	0	0	0	0	0	0	1
17	HAD-Net: A Hierarchical Adversarial Knowledge Distillation Network for Improved Enhanced Tumour Segmentation Without Post-Contrast Images	0	0	0	0	0	0	0	1
18	Guided Filter Regularization for Improved Disentanglement of Shape and Appearance in Diffeomorphic Autoencoders	0	0	0	0	0	0	0	1
19	Learning the Latent Heat Diffusion Process through Structural Brain Network from Longitudinal β-Amyloid Data	0	0	0	0	0	0	0	1
20	MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models	0	0	0	0	0	0	0	1
21	An AI-based Framework for Diagnostic Quality Assessment of Ankle Radiographs	0	0	0	0	0	0	0	1
22	Untangling the Small Intestine in 3D cine-MRI using Deep Stochastic Tracking	0	0	0	0	0	0	0	1
23	Changing the Contrast of Magnetic Resonance Imaging Signals using Deep Learning	0	0	0	0	0	0	0	1
24	Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification	0	0	0	0	0	0	0	1
25	Unseen Disease Detection for Deep Learning Interpretation of Chest X-rays	0	0	0	0	0	0	0	1
26	Embedding-based Instance Segmentation in Microscopy	0	1	0	1	0	0	0	1
27	Explainable Image Quality Analysis of Chest X-Rays	0	0	0	0	0	0	1	1
28	Predicting COVID-19 Lung Infiltrate Progression on Chest Radiographs Using Spatio-temporal LSTM based Encoder-Decoder Network	0	0	0	0	0	0	0	1
29	A Mean-Field Variational Inference Approach to Deep Image Prior for Inverse Problems in Medical Imaging	0	0	0	0	0	0	0	1
30	Benefits of Linear Conditioning for Segmentation using Metadata	0	0	0	0	1	0	0	1
31	Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning	0	0	0	0	0	0	0	1
32	Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features	0	0	1	0	0	0	0	1
33	Feature-based image registration in structured light endoscopy	0	0	0	0	0	0	0	1

This dataset was generated by the notebook I have been working these past couple of days. What you see are keyword matches from MIDL 2021, `0` means no match and `1` indicates at least one.

The most challenging part, which should be obvious from yesterday, was engineering an appropriate crawler for bulk-saving PDFs. Still, I think now I have reached a point where for another venue I will have to only change a few `xpath` selector rules, which is great.

```
#Code snippet for parsing https://proceedings.mlr.press/v143/ crawl
def parse(self, response):
    for article in response.xpath('/html/body/main/div/div[*]'):
        try:
            yield Request(
                url=article.xpath('p[3]/a[2]/@href').get(),
                meta={
                    "title": article.xpath('p[1]/text()').get()
                    },
                callback=self.save_pdf
            )
        except Exception as e:
            print(e)
```

The next step was to use https://github.com/pdfminer/pdfminer.six to extract text from all the PDFs. I used the default `extract_text` function, but it also has a few other modes which can carry layout information by mimicking HTML/XML.

Now for the meat of it. Matching was done on names and keywords I had determined a few days ago. Matches are strict, without edit-distance tolerance or synonyms being considered, and yes, with string comparisons using lower-case text and keywords.

```
mentions = [
    "2017 Data Science Bowl",
    "Kaggle Data Science Bowl 2017",
    "Data Science Bowl 2017",
    "KDSB17",
    "DSB",
]

keywords = [
    "lung cancer",
    "nodule",
    "competition",
    "kaggle dataset",
    "kaggle",
    "deep learning"
]
```

For this batch, I had 0 luck with direct-mentions. If nothing, I did at least get some false-positives for the `DSB` abbreviation; so I know that it's checking.

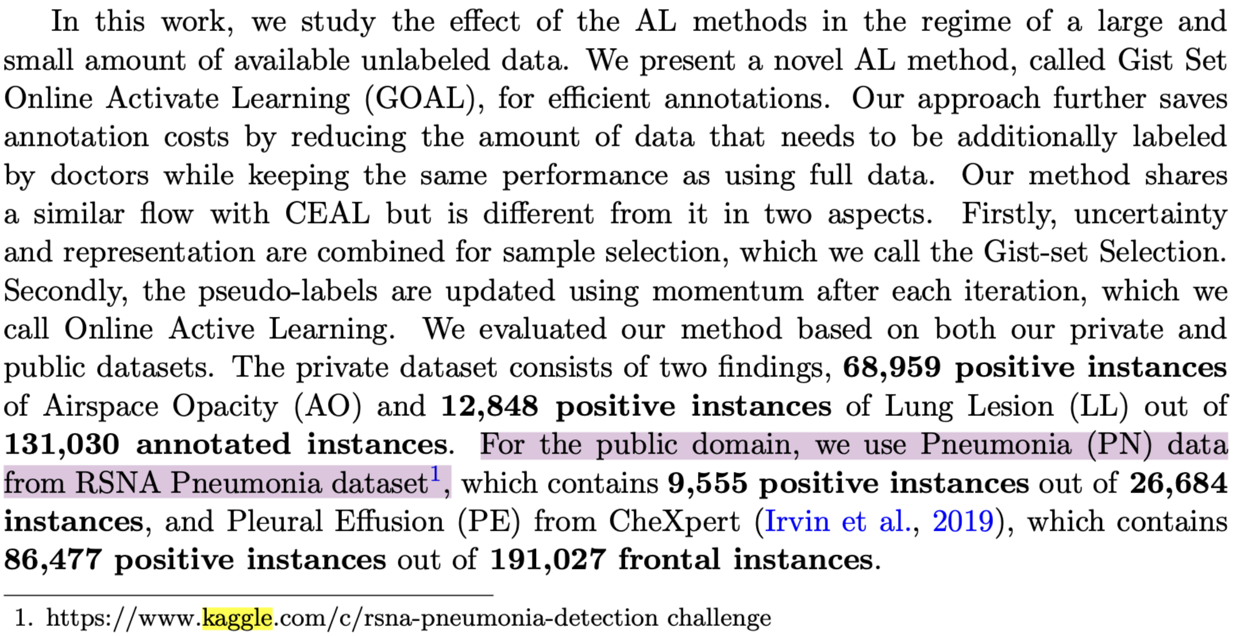
```
Found DSB in Improving MRI-based Knee Disorder Diagnosis with Pyramidal Feature Details : "...wp5ymyj+hedsb5yro2v6..."
Found DSB in Embedding-based Instance Segmentation in Microscopy : "...cision (apdsb, see m..."
```

Keyword-matching went much better. In fact I decided to add a new keyword. `kaggle`, but alone.

```
Found kaggle in GOAL: Gist-set Online Active Learning for Efficient Chest X-ray Image Annotation : "...tps://www.kaggle.com..."
Found kaggle in Embedding-based Instance Segmentation in Microscopy : "...from the kaggle dat..."
```

It is interesting that we can discover certain papers in this this way and I will now show you why:

Let's start with a snippet from the `GOAL` paper



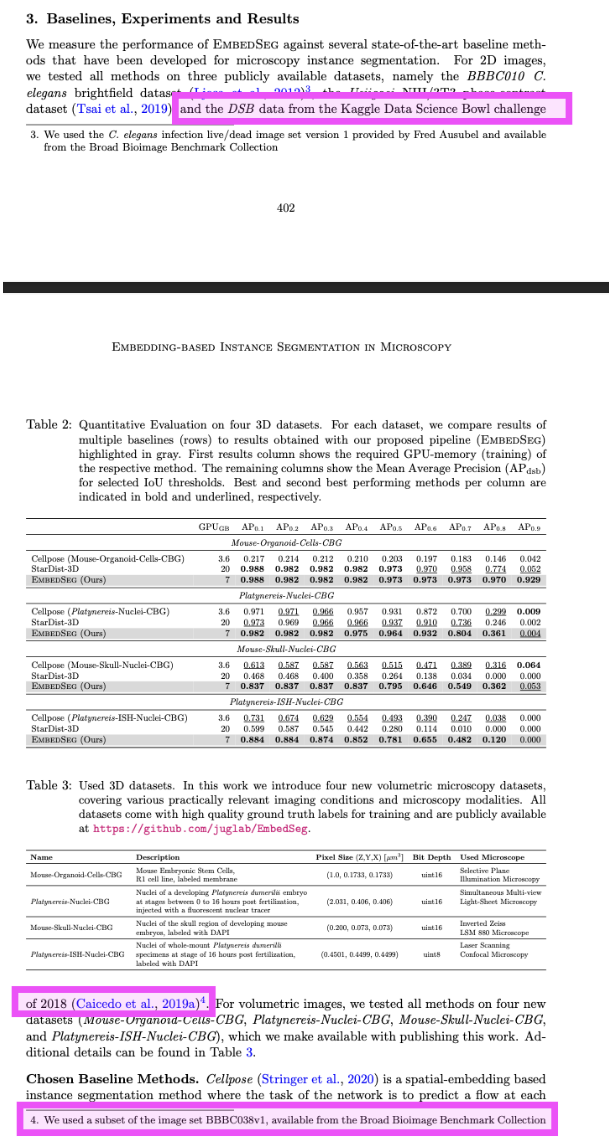
In this work, we study the effect of the AL methods in the regime of a large and small amount of available unlabeled data. We present a novel AL method, called Gist Set Online Activate Learning (GOAL), for efficient annotations. Our approach further saves annotation costs by reducing the amount of data that needs to be additionally labeled by doctors while keeping the same performance as using full data. Our method shares a similar flow with CEAL but is different from it in two aspects. Firstly, uncertainty and representation are combined for sample selection, which we call the Gist-set Selection. Secondly, the pseudo-labels are updated using momentum after each iteration, which we call Online Active Learning. We evaluated our method based on both our private and public datasets. The private dataset consists of two findings, **68,959 positive instances** of Airspace Opacity (AO) and **12,848 positive instances** of Lung Lesion (LL) out of **131,030 annotated instances**. For the public domain, we use Pneumonia (PN) data from RSNA Pneumonia dataset[1], which contains **9,555 positive instances** out of **26,684 instances**, and Pleural Effusion (PE) from CheXpert (Irvin et al., 2019), which contains **86,477 positive instances** out of **191,027 frontal instances**.

[1] https://www.kaggle.com/c/rsna-pneumonia-detection challenge

This is not the `DSB` dataset, instead this dataset originates from an **RSNA Pneumonia Detection Challenge** in 2018. Here are a few remarks:

- This is the only citation to the dataset in the whole paper. Any mention afterwards is abbreviated only as (`PN`). There is nothing in the bibliography either.
- The one linking reference to the dataset is this footnote at the bottom of the page.
- And the worst part is that, unfortunately, the link is broken! It came out as `https://www.kaggle.com/c/rsna-pneumonia-detection challenge` when it should have been `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/`

The second paper, Embedding-based Instance Segmentation in Microscopy does however, deal with a `DSB` dataset. But there's a catch! This `DSB` refers to the **2018** Data Science Bowl, from the year after.



3. Baselines, Experiments and Results

We measure the performance of EmbedSeg against several state-of-the-art baseline methods that have been developed for microscopy instance segmentation. For 2D images, we tested all methods on three publicly available datasets, namely the *BBBC010 C. elegans* brightfield dataset (Ljosa et al., 2012)[3], the *Usiigaci* NIH/3T3 phase contrast dataset (Tsai et al., 2019) and the *DSB* data from the Kaggle Data Science Bowl challenge

[3] We used the *C. elegans* infection live/dead image set version 1 provided by Fred Ausubel and available from the Broad Bioimage Benchmark Collection

402

Embedding-based Instance Segmentation in Microscopy

Table 2: Quantitative Evaluation on four 3D datasets. For each dataset, we compare results of multiple baselines (rows) to results obtained with our proposed pipeline (EmbedSeg) highlighted in gray. First results column shows the required GPU-memory (training) of the respective method. The remaining columns show the Mean Average Precision (AP_{dsb}) for selected IoU thresholds. Best and second best performing methods per column are indicated in bold and underlined, respectively.

	GPU_{GB}	$AP_{0.1}$	$AP_{0.2}$	$AP_{0.3}$	$AP_{0.4}$	$AP_{0.5}$	$AP_{0.6}$	$AP_{0.7}$	$AP_{0.8}$	$AP_{0.9}$
Mouse-Organoid-Cells-CBG										
Cellpose (Mouse-Organoid-Cells-CBG)	3.6	0.217	0.214	0.212	0.210	0.203	0.197	0.183	0.146	0.042
StarDist-3D	20	**0.988**	**0.982**	**0.982**	**0.982**	**0.973**	0.970	0.958	0.774	0.052
EmbedSeg (Ours)	7	**0.988**	**0.982**	**0.982**	**0.982**	**0.973**	**0.973**	**0.973**	**0.970**	**0.929**
Platynereis-Nuclei-CBG										
Cellpose (*Platynereis*-Nuclei-CBG)	3.6	0.971	0.971	0.966	0.957	0.931	0.872	0.700	0.299	**0.009**
StarDist-3D	20	0.973	0.969	0.966	0.966	0.937	0.910	0.736	0.246	0.002
EmbedSeg (Ours)	7	**0.982**	**0.982**	**0.982**	**0.975**	**0.964**	**0.932**	**0.804**	**0.361**	0.004
Mouse-Skull-Nuclei-CBG										
Cellpose (Mouse-Skull-Nuclei-CBG)	3.6	0.613	0.587	0.587	0.563	0.515	0.471	0.389	0.316	**0.064**
StarDist-3D	20	0.468	0.468	0.400	0.358	0.264	0.138	0.034	0.000	0.000
EmbedSeg (Ours)	7	**0.837**	**0.837**	**0.837**	**0.837**	**0.795**	**0.646**	**0.549**	**0.362**	0.053
Platynereis-ISH-Nuclei-CBG										
Cellpose (*Platynereis*-ISH-Nuclei-CBG)	3.6	0.731	0.674	0.629	0.554	0.493	0.390	0.247	0.038	0.000
StarDist-3D	20	0.599	0.587	0.545	0.442	0.280	0.114	0.010	0.000	0.000
EmbedSeg (Ours)	7	**0.884**	**0.884**	**0.874**	**0.852**	**0.781**	**0.655**	**0.482**	**0.120**	0.000

Table 3: Used 3D datasets. In this work we introduce four new volumetric microscopy datasets, covering various practically relevant imaging conditions and microscopy modalities. All datasets come with high quality ground truth labels for training and are publicly available at https://github.com/juglab/EmbedSeg.

Name	**Description**	**Pixel Size** (Z,Y,X) [μm^3]	**Bit Depth**	**Used Microscope**
Mouse-Organoid-Cells-CBG	Mouse Embryonic Stem Cells, R1 cell line, labeled membrane	(1.0, 0.1733, 0.1733)	uint16	Selective Plane Illumination Microscopy
Platynereis-Nuclei-CBG	Nuclei of a developing *Platynereis dumerilii* embryo at stages between 0 to 16 hours post fertilization, injected with a fluorescent nuclear tracer	(2.031, 0.406, 0.406)	uint16	Simultaneous Multi-view Light-Sheet Microscopy
Mouse-Skull-Nuclei-CBG	Nuclei of the skull region of developing mouse embryos, labeled with DAPI	(0.200, 0.073, 0.073)	uint16	Inverted Zeiss LSM 880 Microscope
Platynereis-ISH-Nuclei-CBG	Nuclei of whole-mount *Platynereis dumerilli* specimens at stage of 16 hours post fertilization, labeled with DAPI	(0.4501, 0.4499, 0.4499)	uint8	Laser Scanning Confocal Microscopy

of 2018 (Caicedo et al., 2019a)[4]. For volumetric images, we tested all methods on four new datasets (*Mouse-Organoid-Cells-CBG*, *Platynereis-Nuclei-CBG*, *Mouse-Skull-Nuclei-CBG*, and *Platynereis-ISH-Nuclei-CBG*), which we make available with publishing this work. Additional details can be found in Table 3.

Chosen Baseline Methods. *Cellpose* (Stringer et al., 2020) is a spatial-embedding based instance segmentation method where the task of the network is to predict a flow at each

[4] We used a subset of the image set BBBC038v1, available from the Broad Bioimage Benchmark Collection

Due to the tables you see here, there is a whole page between the mention of the `DSB` data and its reference.

- The 2018 `DSB` is a lot more fortunate than its older sibling, because the authors of the dataset published an article in Nature: `Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl` https://www.nature.com/articles/s41592-019-0612-7
  - It covers what motivated the work, how the data was collected and an assessment of the contest submissions in its aftermath.
- If you are wondering, no, the text didn't capture the connection between `DSB` and the citation to the `2018` challenge well either. But it is at least included in the bibliography this time, even if a reader misses it while reading.

```
and the DSB data from the Kaggle Data Science Bowl challenge

3. We used the C. elegans infection live/dead image set version 1 provided by Fred Ausubel and available

(... many lines later...)

of 2018 (Caicedo et al., 2019a)4. For volumetric images, we tested all methods on four new
datasets (Mouse-Organoid-Cells-CBG, Platynereis-Nuclei-CBG, Mouse-Skull-Nuclei-CBG,
and Platynereis-ISH-Nuclei-CBG), which we make available with publishing this work. Ad-
ditional details can be found in Table 3.
```

- Finally, as it turns out, `apdsb` from the false-positive matches earlier was related to `DSB` after all. It actually stands for `APdsb`, or the Mean Average Precision over the `DSB (2018)` dataset. I guess it's not that easy to write off a match like this as only a coincidence :}

All that is left to do to get a `.csv` of these matches is to bring everything together in a `pandas` dataframe. And that's about it, a fully working pipeline, already (^_^)

Although, it's not without its problems:

- DOIs are missing. They were missing from the proceedings and I don't yet have a strategy to scan for them.
- I'm not filtering, for example, the preface of the proceedings. It was also downloaded and saved as `Proceedings.pdf` as part of the dataset.
- Match frequency isn't taken into consideration. One match is just as good as 100 in this scheme.
    - Also, I have no way to check for homographic false-positives. Hence the `DSB` 2017 vs. 2018 confusion.

I'm not letting those discourage me though, there's still plenty of time to sort these out.

As far as future work goes, searching for `kaggle` was the most revealing thing I came up with. When we have the collection of publications from a particular venue (or set of venues) we can find references to datasets which are missed in most relevance-based indexing systems.

Then perhaps, instead of focussing on `DSB` alone we might try querying the word `kaggle` to extract a family of interrelated datasets across publications.

We should discuss with Veronika during tomorrow's meeting.

Logging off, until next time \(^_^)/.

04-10-2022 : Shovelling

Today wasn't super-productive. I mostly spent my time wrestling with Scrapy. 🙂

03-10-2022 : Keyforge

For starters I'm working with a Jupyter notebook to avoid analysis paralysis. Here are some highlights of what I have been doing recently:

Finding Keywords: I skimmed through the top 4 results on Google Scholar for the query from before on `Data Science Bowl 2017`. Regrettably, they are sorted by `keyword-relevance` instead of total citation count. Although that is still good for our purposes.

Names: Different papers referred to this dataset differently:

- `2017 Data Science Bowl` in `D. Hammack, ‘Forecasting lung cancer diagnoses with deep learning’,` *`Data Science Bowl`* `, pp. 1–6, 2017.`
- `Data Science Bowl 2017` in `K. Kuan` *`et al.`* `, ‘Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017 Challenge’. arXiv, May 26, 2017. doi: 10.48550/arXiv.1705.09435 .`
- `Kaggle Data Science Bowl 2017` in `R. Tekade and K. Rajeswari, ‘Lung Cancer Detection and Classification Using Deep Learning’, in` *`2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)`* `, Pune, India, Aug. 2018, pp. 1–5. doi: 10.1109/ICCUBEA.2018.8697352 .` and `O. Ozdemir, R. L. Russell, and A. A. Berlin, ‘A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans’,` *`IEEE transactions on medical imaging`* `, vol. 39, no. 5, pp. 1419–1429, 2019, doi: 10.1109/TMI.2019.2947595 .`

Abbreviations: Abbreviations are also unfortunately, inconsistent.

- `KDSB17` was used in `Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017 Challenge`
- `DSB` was used in `Forecasting lung cancer diagnoses with deep learning`

Keywords:

- `lung` is too broad, but `Lung Cancer` is always a frequent term.
- `nodule` is an even more frequent term.

- `competition` and `kaggle dataset` are terms used to introduce the dataset.
- `deep learning` is often used as a keyword, but should not be queried alone since it is a very buzzy word.

Future:

- **DOI not yet available in proceedings, I will need a work-around for this! (https://proceedings.mlr.press/v143/)**
- I may consider more quantitive analysis for finding initial keywords on additional datasets.

29-09-2022 : Notes from Bethany

HelloStar Summary

Morning Meeting Takeaways from `HelloStar`, a project by Bethany Chamberlain on thematic intersections of conferences:

- Nvivo Tools, SurveyXact.
- Keyword frequency:
  - “Conference” most common term.
  - MICAI (mis-transcribed as MICA), some cleaning, processing needed
- Surveys:
  - ~34 respondents
  - Not possible to catch up with people in an anonymised-survey. Ask for conference participation.
- “How did you first learn about this conference?”
- “Compare/contrast with other conferences”
  - Big Conferences splitting-up (mitosis style).
  - Inner Circles, Limited Diversity.
- Good Ideas:
  - Send Surveys Early
  - Keep Things Simple (limit drop-out) and **be specific**.
    - Imperatives are better than being kind.
      - “Can you describe this?”
        - “No.”
  - Follow-up Interviews can yield additional information.
    - Alternatively, Focus Groups.

Zotero

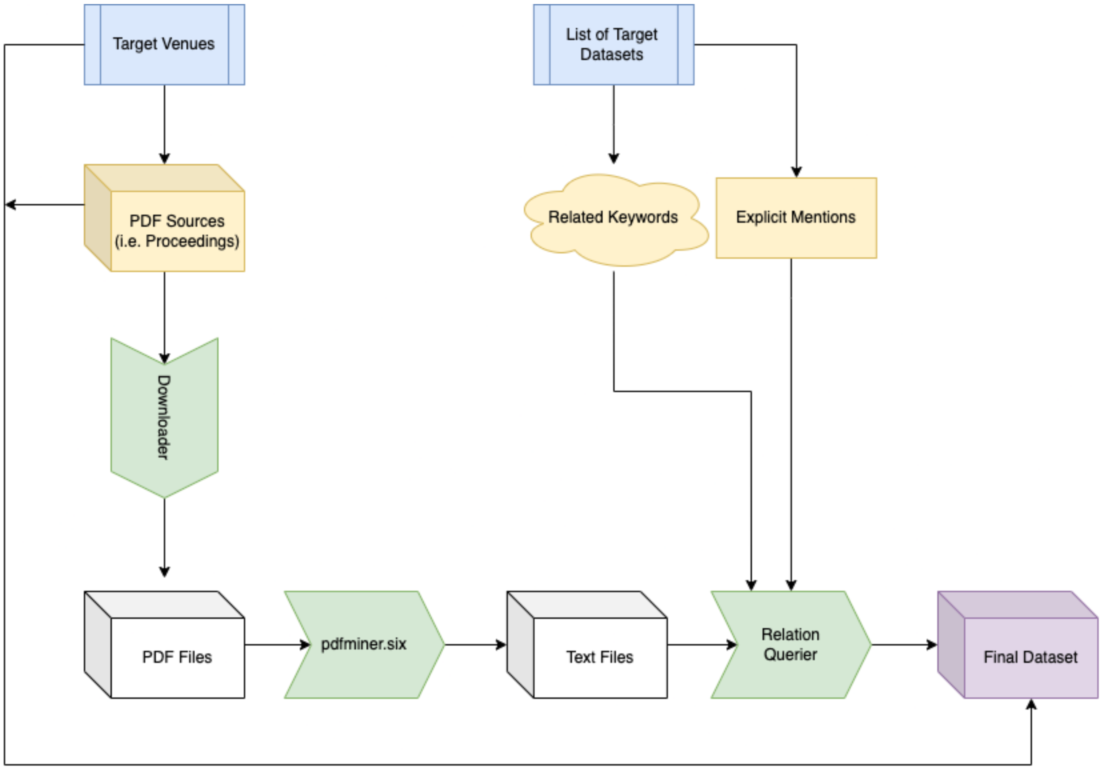
Some more notes from Bethany’s talk:

- “Was competing with Notion, but Zotero was all I really needed.”
  - Zotero also has nice Markdown support for notes!
    - Zotero Better Notes (requires cloud :( )
- Cloud Syncing? Worth it imo
  - Not imo :)
- Collaboration
  - First time this semester with MSc. Students :)
  - I have my library on Notion (Zotero Library link goes here)
  - Updates are propagated fast

  - Trial and Error
  - “Retracted Item” warning for anything deleted abruptly.
- Add-Ons <3
  - **DOI Manager**
  - **Storage Scanner for Zotero: Duplicates**
  - **Zotero PDF Preview**
  - Zotero OCR
  - Zotero Better Notes

28-09-2022 : Engineering

Today, I can finally piece together a proper pipeline to build our project dataset.



Blueprint: Target `Venue`s and Target `Dataset`s

Input Parameters: Since `venue`s are not universally accessible, the input for `PDF Sources` is comprised of `URL`s pointing to proceedings and similar collections to be crawled or queried. Likewise, there is (yet) no obvious way to derive the related `keyword`s and the common `name` used for a target dataset without direct investigation. The target `dataset`s are treated as blueprints, and the corresponding input is modelled accordingly.

Scripts/Code:

- Downloader: Downloads `publications` as `PDF` Files, and organises them by `venue` label.
- `pdfminer.six`: Is called to process `PDF`s into `text` files.
- Relation Querier: Queries `text` files for mentions of the input `keyword`s on or the common name of the target `dataset`.

Intermediates:

- PDF Files: Labeled by venue, with DOI also extracted if available.

- Text Files: Labeled by venue, corresponding to PDF Files.

Output: Final Dataset

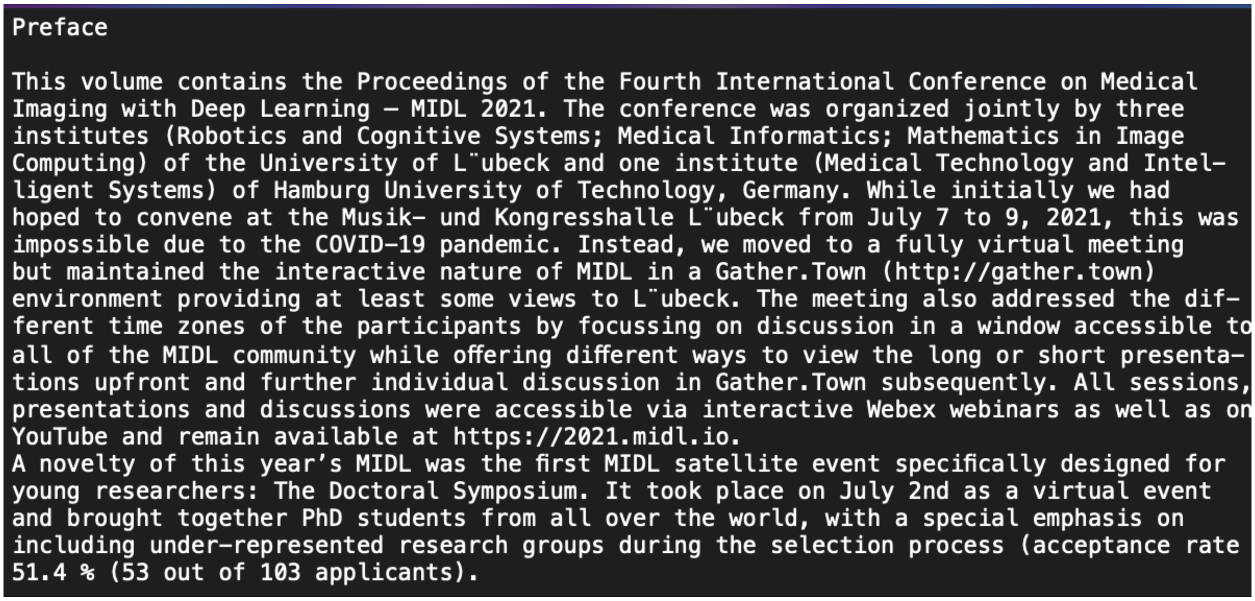
- CSV
  - Venue
  - Publication Title
  - DOI
  - Boolean-Variables Indicating `keyword` and `dataset`-name matches.

27-09-2022 : PDF TExtraction

Alright, how do we avoid opening and control-F'ing for keywords in each PDF? No, not `OCR`; there are PDF decoders that can extract text after all

I had high hopes for https://github.com/kingaling/pydf2json. But it's outdated (Python 2), barely maintained and perhaps too verbose. I spent a whole evening trying to make it work, and even when it did produce some output, mixed with literal and hexadecimal strings, it required further post-processing.

So instead, I switched to pdfminer.six which offers formatted text output, as well as pseudo-XML.



```
Preface

This volume contains the Proceedings of the Fourth International Conference on Medical
Imaging with Deep Learning – MIDL 2021. The conference was organized jointly by three
institutes (Robotics and Cognitive Systems; Medical Informatics; Mathematics in Image
Computing) of the University of L¨ubeck and one institute (Medical Technology and Intel-
ligent Systems) of Hamburg University of Technology, Germany. While initially we had
hoped to convene at the Musik- und Kongresshalle L¨ubeck from July 7 to 9, 2021, this was
impossible due to the COVID-19 pandemic. Instead, we moved to a fully virtual meeting
but maintained the interactive nature of MIDL in a Gather.Town (http://gather.town)
environment providing at least some views to L¨ubeck. The meeting also addressed the dif-
ferent time zones of the participants by focussing on discussion in a window accessible to
all of the MIDL community while offering different ways to view the long or short presenta-
tions upfront and further individual discussion in Gather.Town subsequently. All sessions,
presentations and discussions were accessible via interactive Webex webinars as well as on
YouTube and remain available at https://2021.midl.io.
A novelty of this year's MIDL was the first MIDL satellite event specifically designed for
young researchers: The Doctoral Symposium. It took place on July 2nd as a virtual event
and brought together PhD students from all over the world, with a special emphasis on
including under-represented research groups during the selection process (acceptance rate
51.4 % (53 out of 103 applicants).
```

This is much better.

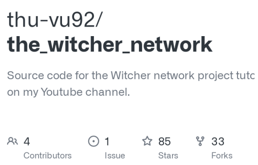
24-09-2022 : Toss a coin to your Witcher!

While doing some work-related research, I stumbled upon **Thu Vu**'s very recent `Witcher Network` project. Good example for basic NLP (named-entity recognition, feature-engineering) and network visualisation.

GitHub - thu-vu92/the_witcher_network: Source code for the Witcher network project tutorial on my Youtube channel.

You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

https://github.com/thu-vu92/the_witcher_network



thu-vu92/
the_witcher_network

Source code for the Witcher network project tutc
on my Youtube channel.

https://www.youtube.com/watch?v=RuNolAh_4bU&t=607s

https://www.youtube.com/watch?v=fAHkJ_Dhr50

22-09-2022 : The Hidden Popularity of Machine Learning Challenges

The biggest take-away from today’s meeting for me was getting to rename this project, which I feel has several layers of depth. **The Hidden Popularity of Machine Learning Challenges**

I chose this name because I see that the project is moving away from pure Machine Learning to responsible machine learning. The previous name ***Predicting Popularity of Machine Learning Challenges*** had a connotation of predicting popularity as the output of some kind of model. But as I have seen over the past few days, we don’t yet have the resources to get there. Through citation negligence, identifier registration and the other hassles of the academic milieu, it is already challenging as-is to uncover references to Machine Learning Challenges. I feel that is the bulk of where I am working, and any inferential statistics will be the icing on the cake.

Our over-arching project is code-named `Public Datasets`, but I have chosen to stick to `Machine Learning Challenge` datasets, specifically. So I kept that part of the name in tact. I now see this as very much a hotspot of unsolved problems and that draws me closer, giving me reason to focus.

The next big challenge will be building a pipeline to bulk-download PDFs and extract text for processing. That will spare me of having to manually annotate everything and give a very valuable potential for scaling up.

However Veronika as my supervisor also cautioned me of two things:

- It might be a good idea to refine keywords, based on what we can gleam from papers mentioning `Data Science Bowl 2017` explicitly:

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=%22data+science+bowl%22+%2B+lung&btnG=

- To avoid DOI-hunting, it’s better to work with finer curated proceedings. I will stick to `MIDL` for now, but I will be testing my PDF-extractor and querier using `MIDL 2021` papers instead.

Proceedings of Machine Learning Research

Proceedings of the Fourth Conference on Medical Imaging with Deep Learning Held in Lübeck, Germany on 07-09 July 2021 Published as Volume 143 by the Proceedings of Machine Learning Research on 25 August 2021. Volume Edited by: Mattias Heinrich Qi Dou Marleen de Bruijne Jan Lellmann Alexander Schläfer Floris Ernst Series Editors: Neil D.

https://proceedings.mlr.press/v143/

20-09-2022 : Semantic Searching

After past blunders with brute-force querying, I spent today mostly catching up on semantic-searching. Flexible queries which can match contextual synonyms. There are large-scale production tools like Solr, but those seem overkill for now. Still interesting to learn about.

How to build semantic search for a given domain

There is a problem we are trying to solve where we want to do a semantic search on our set of data, i.e we have a domain-specific data (example: sentences talking about automobiles) Our data is just a bunch of sentences and what we want is to give a phrase

https://stackoverflow.com/a/60312352/2089784 https://www.youtube.com/watch?v=4fMZnunTRF8

- Likely, I will be working with `gensim`

How to do string semantic matching using gensim in Python?

You can use gensim library to implement MatchSemantic and write code like this as a function (see full code in here): pip install numpy pip install gensim first of all, we must implement the requirements from re import sub import numpy as np from gensim.utils import simple_preprocess import gensim.downloader as api from

https://stackoverflow.com/a/71828372/2089784

- But `Hugging Face` is making transformers really viable, so maybe it's better to get with the times.

https://www.youtube.com/watch?v=OATCgQtNX2o

So these would allow me more versatility in searching for contest or topic mentions, but I still need to work out my pipeline for downloading articles and converting them to searchable documents.

Logging off, until next time \(^_^)/.

19-09-2022 : Citation vs. Attribution

As a refresher, the first public dataset I'm searching for are mentions of the Lung Image dataset from Kaggle Data Science Bowl '17. To that end, I'm querying for the terms `Data Science Bowl 2017`, `Nodule`, `Lung`, `Lung Tissue` and `Lung Imaging` in either the body, keywords or title of my collected papers. Intuitively this sounded like a good idea; until, of course, some things went wrong.

For example, *MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images (Graham et al., 2018)* and *Evaluating Reinforcement Learning Agents for Anatomical Landmark Detection (Alansary et al., 2018)* cite work on `Lung Cancer` and `Lymph Node` detection in related work, but with no explicit mention in the body. They are presented as applications of machine learning, or as one paper put it: `computerized techniques`. To quote from the *MILD-Net* paper which drops a lengthy citation in the middle of the introduction:

> 1. Introduction
>
> ...
>
> Computerized techniques play a significant role in automated digitalized histology image analysis, with applications to various tasks including but limited to nuclei detection and segmentation (Graham, Rajpoot, 2018, Chen, Qi, Yu, Dou, Qin, Heng, 2017, Sirinukunwattana, Raza, Tsang, Snead, Cree, Rajpoot, 2016), mitosis detection (Cireşan, Giusti, Gambardella, Schmidhuber, 2013, Chen, Dou, Wang, Qin, Heng, et al., 2016, Veta, Van Diest, Willems, Wang, Madabhushi, Cruz-Roa, Gonzalez, Larsen, Vestergaard, Dahl, et al., 2015, Albarqouni, Baur, Achilles, Belagiannis, Demirci, Navab, 2016), tumor segmentation (Qaiser et al., 2017), image retrieval (Sapkota, Shi, Xing, Yang, 2018, Shi, Xing, Xu, Xie, Su, Yang, 2017), cancer type classification (Graham, Shaban, Qaiser, Khurram, Rajpoot, 2018, Kong, Wang, Li, Song, Zhang, 2017, Bejnordi, Veta, Van Diest, Van Ginneken, Karssemeijer, Litjens, Van Der Laak, Hermsen, Manson, Balkenhol, et al., 2017, Lin, Chen, Dou, Wang, Qin, Heng, 2018, Qaiser, Mukherjee, Reddy Pb, Munugoti, Tallam, Pitkäaho, Lehtimäki, Naughton, Berseth, Pedraza, et al., 2018), etc.
>
> ...

> Graham, Shaban, Qaiser, Khurram, Rajpoot, 2018
> S. Graham, M. Shaban, T. Qaiser, S.A. Khurram, N. Rajpoot
> Classification of lung cancer histology images using patch-level summary statistics
> Medical Imaging 2018: Digital Pathology, 10581, International Society for Optics and Photonics (2018), p. 1058119

Another paper, with some shared authors, titled *Attention U-Net: Learning Where to Look for the Pancreas (Oktay et ak., 2018)* is a lot more explicit in connecting `lung`-related medical imaging tasks to the paper:

> 1. Introduction
>
> Automated medical image segmentation has been extensively studied in the image analysis community due to the fact that manual, dense labelling of large amounts of medical images is a tedious and error-prone task. Accurate and reliable solutions are desired to increase clinical work flow efficiency and support decision making through fast and automatic extraction of quantitative measurements. With the advent of convolutional neural networks (CNNs), near-radiologist level performance can be achieved in automated medical image analysis tasks including cardiac MR segmentation [3] and cancerous lung nodule detection [17].
> ...
>
> [17] Liao, F., Liang, M., Li, Z., Hu, X., Song, S.: Evaluate the malignancy of pulmonary nodules using the 3D deep leaky noisy-or network. arXiv preprint arXiv:1711.08324 (2017)

On my part, I was only able to mark the *Attention U-Net* paper as relating to `lung`s and `nodule` detection, even if it may be argued that the *MILD-Net* and *Reinforcement Learning Agents* papers also connected to this work. None of these are papers on lung imaging, but only derive some influence therefrom. Thus there were no experiments involving lung images here, at all.

Moving onto papers which **do** actually deal with lung imaging, even then it's not always clear at what granularity the work is relevant. The paper *Temporal Interpolation via Motion Field Prediction (Zhang et al., 2018)* refers to `Lung Tissue` as `Lung Structure` and so I discovered that my precise keyword-search for `Lung Tissue` was missing something which could be considered relevant. Perhaps, I ought to refine my method to be able to catch synonyms or similar terms.

We tend to hold citation in a higher regard to attribution, that is to say the precise reference to related works is considered more valuable than the concrete explanation of the connection to said related works. While this practice has the advantage of making it easier to find the original work, it doesn't tell where and what for. An informal quote exchanged between colleagues can, surprisingly, reveal a whole lot more.

These are all things for future authors to consdider, but as to me I'm going to need a more flexible search rule if I intend to get anywhere.

Logging off, until next time \(^_^)/.

18-09-2022 : On Identifiers

Last week I collected oral papers from MIDL 2018* on openreview.net. This was a conference that took place right after Data Science Bowl '17 so it seemed like a good source to observe immediate feedback.

* (Medical Imaging with Deep Learning)

I couldn't help but notice that many of these papers were missing their DOI. A minority of authors had access to arXiv so that gave them an easy DOI for preprints. Other authors had to wait between 1 or 3 years to get published in some journal, usually owned by Elsevier, to receive a proper DOI. As I myself am on the fence of whether or not I want to continue onto a PhD, this was a nice reminder that you should never expect immediate recognition for your work. Even when deserved, it has to be earned.

For example, let's take the paper: *Recurrent Inference Machines for Accelerated MRI Reconstruction* by Lønning et al. This version of the paper from 2018 has no DOI. But it does have some other identifiers. It came out of the Spinoza Centre for Neuroimaging in the Netherlands, a country which has several national identifers. Thus it is possible to identify the paper by NBN or Handle. Another version of *The Recurrent Inference Machines* paper did eventually receive a DOI, titled *Recurrent inference machines for reconstructing heterogeneous MRI data* in the journal *Medical Image Analysis*. As you can imagine, this did not happen until 2019 and *Medical Image Analysis* is also published by Elsevier.

But of course, I only need identifers for future reference, the papers themselves are from an open-access conference after all. Unfortunately, `openreview.net` is no `wikipedia.org`, it does not usher the same level of trust and permanence. Therefore I feel identifying papers by link alone would be a liability for reproducability.

For the time being, I have only marked those papers whose identifiers I was not able to find as `n/a`. I must admit that I find it quite funny how while I am trying to examine the negotiation between academia and the web, that I also have to put-up with the internal struggles within Academia itself as well. Now I'm left wondering if I should stick to listing DOI-only or bite the bullet and have links as an alternative identifier.

Logging off, until next time (\^_^/).

The Story so Far:

My initial approach has been to work backwards: To first collect literature and *then to scan* for mentions of corresponding datasets. This is solely motivated by the fact that there are a lot more tools to query literature.

The first dataset of interest comes from the Kaggle Data Science Bowl '17. This was a lung-cancer tissue detection challenge with annotated data provided by The National Lung Screening Trial, Copenhagen University Hospital and others. Arguably, the challenge and corresponding dataset inspired a new interest in lung cancer research; see (Varoquaux & Cheplygina, 2022).

Read more here:

Misc. Notes:

Trello

Your browser was unable to load all of Trello's resources. They may have been blocked by your firewall, proxy or browser configuration.Press Ctrl+F5 or Ctrl+Shift+R to have your browser try again and if that doesn't work, check out our troubleshooting guide .

https://trello.com/b/osa62908/predicting-popularity-of-machine-learning-challenges