

Human-Centered AI Research Seminar

Module 3: The Mechanics: Management of Research

Andreas Holzinger
Human-Centered AI (Holzinger Group)
Institute for Medical Informatics/Statistics, Medical University Graz, Austria
and
Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada



@aholzin #KandinskyPatterns

Course Homepage: <https://human-centered.ai/hcai-research-seminar-2020>

- In Module 2 we discussed questions including *“What is science?”*, *“Why contributing to the international scientific community?”* and some *methodological* issues, now
- in Module 3 we discuss questions including *“How to contribute to the international scientific community?”* and learn the basic mechanics of science, the *“know-how”*,
- Of course always from our human-centered AI and ethical responsible machine learning perspective

This is the version for
printing and reading.
The lecture version is
didactically different.

friday 14 October lecture on variational inference.

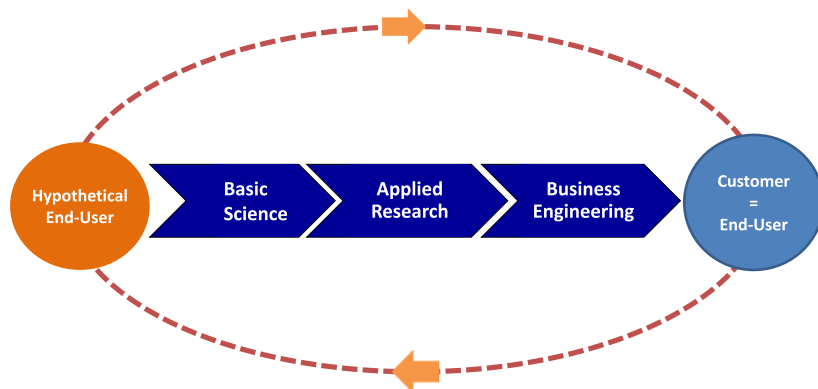
Final Project

In the second half of the course, you will complete a project. The ideal outcome of this project would be a paper that could be submitted to a top-tier machine learning conference such as NIPS, ICML, UAI, AISTATS, or KDD. There are different ways to approach this project, which are discussed in a more comprehensive document that is available from the course website under the Files tab. There are four separate components of the project:

... to contribute to the
international scientific
community

01 Successful Management of Research & Development

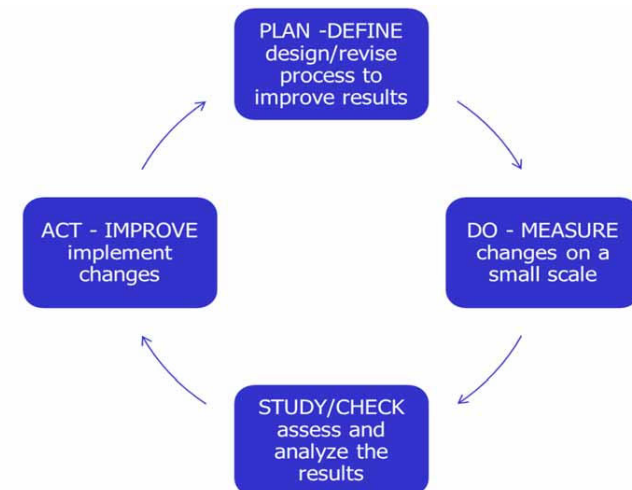
Science is testing crazy ideas –
Engineering is putting these ideas into Business



Holzinger, A. 2011. Successful Management of Research and Development, Norderstedt: BoD.

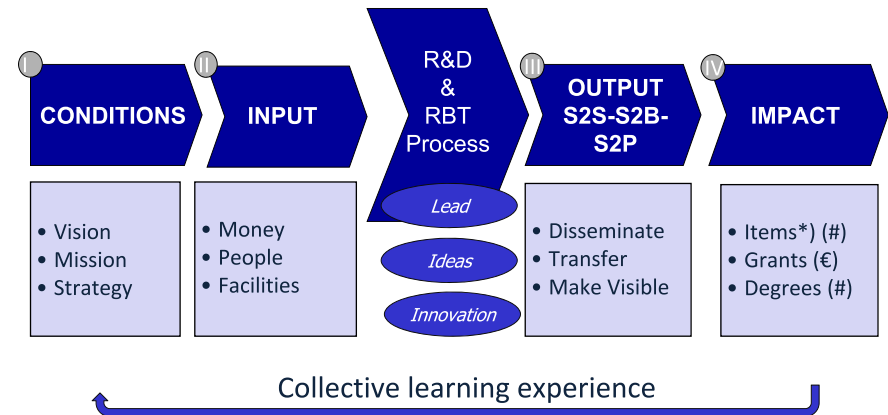
- Research & Development (R&D) play a central role in AI in order to retain and to expand the sustained competitiveness of our societies. Successful management of R&D ensures effectiveness, efficiency, and the quality of all necessary factors and processes in terms of enabling sustained progress in AI.

"If you ask what real knowledge is, I answer, that which enables action"
Hermann von Helmholtz (1821-1894)



Adri Platje, Harald Seidel & Sipke Wadman 1994. Project and portfolio planning cycle: project-based management for the multiproject challenge. International Journal of Project Management, 12, (2), 100-106.

02 Workflows for a Research Group



*) patents, publications, ... any countable output to the international research community

Holzinger, A. 2011. *Successful Management of Research and Development*, Norderstedt, BoD.

- ad I) PRECONDITIONS
- **1. Vision** (is a *declaration* of the goal you *want* to achieve and is necessary for successful research)
- **2. Mission** (*statements* which guides the actions towards the goal and provide a *visible direction*)
- **3. Strategy** (the overall *plan of action* to get closer to the goal)
- If a strategy has more than one page - than it is none!

- ad II) INPUT
- **1. Money** (money is not everything, but everything is nothing without money)
- **2. People** (effective team building is most crucial for success, it is you ...)
- **3. Facilities** (coffee-machines, infrastructure, equipment, tools*, ...)

*) a fool with a tool is still a fool ...

- ad III) OUTPUT
- **1. Dissemination** (science-to-science: deliverables leading to measurable items, e.g. algorithms, solutions, prototypes, tools, data, patents, papers, ... contributes to and helps the international research community!)
- **2. Knowledge Transfer** (teaching to group, teaching to faculty, science-to-industry, ...)
- **3. Visibility** (community building, networking, event organization (e.g. workshop, sessions, conferences), science-to-public, ...)

- ad IV) IMPACT
- The impact can be determined by everything which is measurable and countable, and which can be enlisted and entered into a science balance (“Wissensbilanz”):
- **1. Items** (# patents, publications in journals, proceedings, book chapters, ...)
- **2. Grants** (EUR amount of raised funding, sponsorship, ...)
- **3. Degrees** (# supervised Bachelors, Masters, PhDs, ...)

03 Measurable Output: Publications “papers”

What is a “paper”?

- is a **message** to the international research community (you “have something to say” ...)
- is written in the scientific language today - **English** (Latin in medieval times and Greek in ancient times)
- reports something **of value** for other researchers
- should be useful, and what other researchers can use, will be **referenced** (brings you citations)
- is subject to **peer review**
- has a **specific form** (style and format)
- appears within a scientific **journal** or in **conference proceedings** (e.g. Springer LNCS/LNAI)



Note: Should be included in the DBLP: <https://dblp.uni-trier.de/>

Journals should be included in SCI: <https://webofknowledge.com>

A subjective overview can be found: <https://human-centered.ai/ai-machine-learning-related-journals>

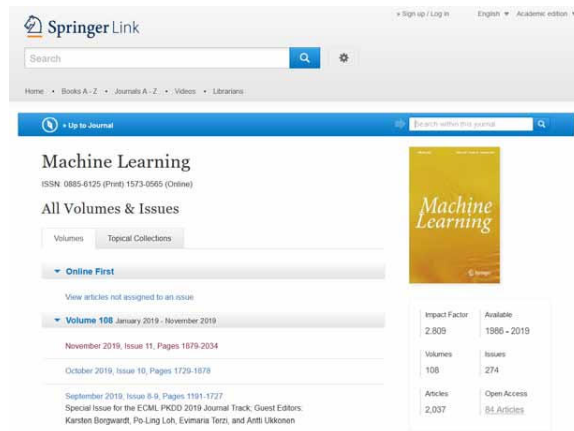
Not to be confused with Student Textbooks or Scientific Monographs (“Books”):



Machine Learning is an international forum for research on computational approaches to learning. The journal publishes articles reporting substantive results on a wide range of learning methods applied to a variety of learning problems.

The journal features papers that describe research on problems and methods, applications research, and issues of research methodology. Papers making claims about learning problems or methods provide solid support via empirical studies, theoretical analysis, or comparison to psychological phenomena. Applications papers show how to apply learning methods to solve important applications problems. Research methodology papers improve how machine learning research is conducted.

<https://www.springer.com/journal/10994>



<https://link.springer.com/journal/volumesAndIssues/10994>

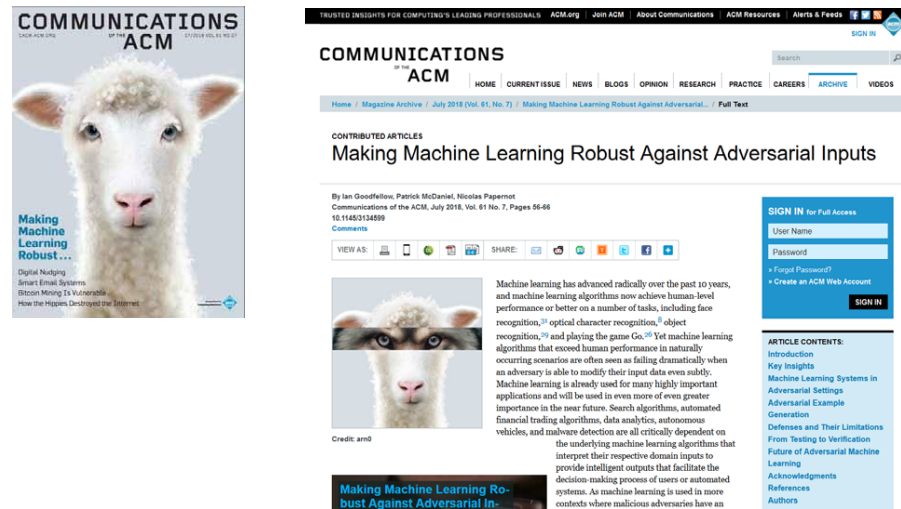


<https://link.springer.com/book/10.1007/978-3-319-99740-7>

Lecture Notes in Computer Science (LNCS)

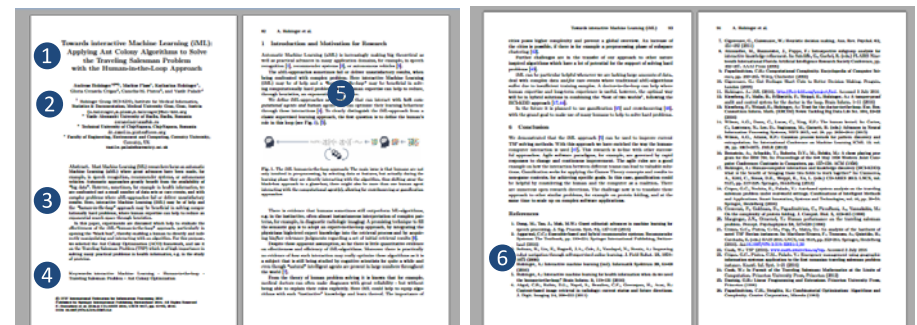
This distinguished conference proceedings series publishes the latest research developments in all areas of computer science. Together with its subseries LNAI & LNBI, LNCS volumes are indexed in the Conference Proceedings Citation Index (CPCI), part of Clarivate Analytics' Web of Science; Scopus; EI Engineering Index; Google Scholar; DBLP; etc.

<https://www.springer.com/gp/computer-science/lncs>



<https://cacm.acm.org/magazines/2018/7/229030-making-machine-learning-robust-against-adversarial-inputs/fulltext>

04 How do I read a paper



- 1 = Paper title
- 2 = Authors with Affiliations
- 3 = Abstract
- 4 = Keywords
- 5 = Content (formally divided into: 1) Introduction and

Motivation for Research > 2) Background and Related Work, 3) Experimental method, setting, results, 4) Discussion, 5) Future Work, 6) Conclusion)

- 6 = References

contributed articles

Such inputs distort how machine-learning-based systems are able to function in the world as it is.

BY IAN GOODFELLOW, PATRICK MCDANIEL, AND NICOLAS PAPERNOT

Making Machine Learning Robust Against Adversarial Inputs

MACHINE LEARNING HAS advanced radically over the past 10 years, and machine learning algorithms now achieve human-level performance or better on a number of tasks, including face recognition,¹⁰ optical character recognition,¹¹ object recognition,¹² and playing the game Go.¹³ Yet machine learning algorithms that exceed human performance in naturally occurring scenarios are often seen as failing dramatically when an adversary is able to modify their input data even subtly. Machine learning is already used for many highly important applications and will be used in even more of even greater importance in the near future. Search algorithms, automated financial trading algorithms, data analytics, autonomous vehicles, and malware detection are all critically dependent on the underlying machine learning algorithms that interpret their respective domain inputs to provide intelligent outputs that facilitate the decision-making process of users or automated

systems. As machine learning is used in more contexts where malicious adversaries have an incentive to interfere with the operation of a given machine learning system, it is increasingly important to provide protections, or "robustness guarantees," against adversarial manipulation.

The modern generation of machine learning services is a result of nearly 50 years of research and development in artificial intelligence—the study of computational algorithms and systems that reason about their environment to make predictions.¹⁴ A subfield of artificial intelligence, most modern machine learning, as used in production, can be essentially be understood as applied function approximation, when there is some mapping from an input x to an output y that is difficult for a programmer to describe through explicit code, a machine learning algorithm can learn an approximation of the mapping by analyzing a dataset containing several examples of inputs and their corresponding outputs. The learning proceeds by defining a "model," a parametric function describing the mapping from inputs to outputs. Google's image-classification system, Inception, has been trained with millions of labeled images.¹⁵ It can classify images as cats, dogs, airplanes, boats, or more complex concepts on par or improving on human recognition.

Key Insights

- Machine learning has traditionally been developed following the assumption that training and evaluation of the model, while useful for designing effective algorithms, this implicitly rules out the possibility that an adversary could alter the distribution at either training time or test time.
- In the context of adversarial inputs at test time, the underlying machine learning model for the many attacks that have been
- To end the arms race between attackers and defenders, we suggest building more tools for verifying machine learning models, unlike current testing practices, this could help industry eventually gain a fundamental advantage.

58 COMMUNICATIONS OF THE ACM • JULY 2018 • VOL. 61 • NO. 7

The best way of reading a paper is to review a paper!

- Goals: What are **research goals** and what are machine learning tasks?
- Description: Is the description adequately detailed for others to **replicate** the work? Is it clearly written in good style and does it include examples? Papers describing systems should clearly describe the contributions or the principles underlying the system. Papers describing theoretical results should also discuss their practical utility.
- Evaluation: Do the authors evaluate their work in an adequate way (theoretically and/or empirically)? Are all claims clearly articulated and supported either by **empirical experiments or theoretical analyses**? If appropriate, have the authors **implemented** their work and demonstrated its utility on a significant problem?
- Significance: Does the paper constitute a significant, technically correct **contribution to the field** that is appropriate for machine learning? Is it sufficiently different from prior published work (by the author or others) to **merit a new publication**? Is it clear how the work advances the current state of understanding, and **why the advance matters**?
- Related Work and Discussion: Are **strengths and limitations** and generality of the research adequately discussed, in particular in relation to related work? Do the authors clearly acknowledge and identify the contributions of their predecessors?
- Clarity: Is it written in a way such that an interested reader with a background in machine learning, but no special knowledge of the paper's subject, could understand and appreciate the paper's results? In particular, is it written in a clear, **readable** style, with good grammar and few (if any) typographical errors?
- Are the **goals and contributions** of the work clearly and correctly stated? Are the problem description, approach and evaluation adequately detailed for others to replicate the work?
- If the paper introduces new terminology or techniques, does it explain **why current terminology or techniques are insufficient**? Does it include **examples**?
- Recommendation: accept (as is), conditional accept (given minor or major revisions), reject with encouragement to revise and resubmit, or definitely reject. If you suggest conditional accept, always provide a precise list of changes that can easily be checked upon resubmission, i.e. that the authors can write a clear rebuttal letter stating on how they reacted on the specific reviewer comments.

Source: <http://www.jmlr.org/reviewer-guide.html>

- 1) Download the review template (link below)
- 2) Fill out the template accordingly
- 3) Rename it, having the review number and acronym in the file name
- 4) Save it as pdf

<https://human-centered.ai/wordpress/wp-content/uploads/2019/10/REVIEW-TEMPLATE-2020-XXXX.doc>

```

***BEGIN of Review***
¶
Title of the Paper:¶
¶
Please describe briefly with your own words what this paper is about:¶
¶
This paper reports on: ... (do not evaluate at this point, just describe with your own words)¶
¶
1) Originality: Does the paper contain significant content to justify publication? What are novel aspects? Did you check for plagiarism, e.g. with a quick Google search?¶
¶
Novel aspects include the topic: ...¶
¶
2) Related Work: Is there enough background and relevant related work? Are any relevant references missing? Please provide recommendations.¶
¶
The following important related papers are missing: ...¶
¶
3) Methodology: Is the paper's argument built on an appropriate base of theory and concepts? Are the methods used appropriately described?¶
¶
The methods are: ...¶
¶
4) Results: Are the results presented clearly and appropriately? Do the conclusions adequately tie together the other elements of the paper?¶
¶
The results are: ...¶
¶
5) Qualitative Evaluation: Is the paper well-written? Is it clear, readable and comprehensive? Sentence structure, acronym explanation, typos, etc. ok?¶
¶
The paper is: ...¶
¶
6) Quantitative Evaluation: Given that the worst paper you have ever read receives 0 and the best paper ever receives 100 points – how many points would you assign to this paper: XX (0 ... 100)¶
¶
FINAL RECOMMENDATION¶
A=Accept – B=Minor Revision – C=Major Revision – D=Reject¶
¶
In case of A, B, or C – please outline how the authors can improve their paper: What should the authors do? What should they expand/remove/etc.? What should they improve? What would you like to read?¶
(Use additional space as you need it)¶
¶
¶
¶
¶
¶
***END of review***

```

HCAI Review Form 1/2019 Thank you very much for your time and effort – your help is most appreciated!

- Stop Reading! Start Writing!
- Read along when writing!
- SQ3R-Method:
 - Survey (read title, abstract, conclusion, subheadings)
 - Question (what are the major insights of this paper?)
 - Read: with regard to the question above
 - Recite: summarize with your own words
 - Review: Try to reflect the major insights of the paper
- Do not waste time! Be economic! It is simply impossible to read everything and all!

05 How to find (relevant) Papers?

Google Scholar explainable ai

Artikel Ungenau 36 900 Ergebnisse (0,06 Sek.)

Beliebige Zeit
Seit 2019
Seit 2018
Seit 2015
Zeitraum wählen...

Nach Relevanz sortieren
Nach Datum sortieren

Beliebige Sprache
Seiten auf Deutsch

☒ Patente einschließen
☒ Zitate einschließen
☐ Alert erstellen

[PDF] Explainable artificial intelligence (xai)
D Gunning - Defense Advanced Research Projects Agency (DARPA ... 2017 - darpa.mil
... 2 The Need for **Explainable AI** User • Why did you do that ... **Explainable AI** will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners **AI System** http://listverse.com/ ©2007–2017 Listverse Ltd ...
☆ 99 Zitiert von: 297 Ähnliche Artikel Alle 6 Versionen In EndNote importieren

What do we need to build explainable AI systems for the medical domain?
A Holzinger, C Biemann, CS Pattichis ... - arXiv preprint arXiv ... 2017 - arxiv.org
Artificial intelligence (AI) generally and machine learning (ML) specifically demonstrate impressive practical success in many different application domains, eg in autonomous driving, speech recognition, or recommender systems. Deep learning approaches, trained ...
☆ 99 Zitiert von: 86 Ähnliche Artikel Alle 4 Versionen In EndNote importieren

Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences
T Miller, P Howe, L Sonenberg - arXiv preprint arXiv:1712.00547, 2017 - arxiv.org
In his seminal book *The Inmates are Running the Asylum: Why High-Tech Products Drive Us Crazy And How To Restore The Sanity* [2004, Sams Indianapolis, IN, USA], Alan Cooper argues that a major reason why software is often poorly designed (from a user perspective) ...
☆ 99 Zitiert von: 73 Ähnliche Artikel Alle 6 Versionen In EndNote importieren

https://scholar.google.at/scholar?hl=de&as_sdt=0%2C5&q=explainable+ai&btnG=

Semantic Scholar All Fields Kandsky

Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology
Andreas Holzinger, Bernd Malle, +4 authors Kurt Zatlowski - Published in *ArXiv* 2017

18 Twitter Mentions

Digital pathology is not only one of the most promising fields of diagnostic medicine, but at the same time a hot topic for fundamental research. Digital pathology is not just the transfer of histopathological slides into digital representations. The combination of different data sources (images, patient records, and tomics data) together with current advances in artificial intelligence/machine learning enable to make novel information accessible and quantifiable to a human expert, which is not... CONTINUE READING

[VIEW PDF ON ARXIV](#) [SAVE TO LIBRARY](#) [CREATE ALERT](#) [CITE](#)

ABSTRACT FIGURES AND TOPICS 13 CITATIONS 66 REFERENCES RELATED PAPERS

Figures and Topics from this paper.

Figures

Explore Further: Topics Discussed in This Paper

- Machine learning
- Artificial intelligence
- Omics
- Laplacian matrix
- Deep learning
- Debugging

<https://www.semanticscholar.org/paper/Towards-the-Augmented-Pathologist%3A-Challenges-of-in-Holzinger-Malle/739c634cd54b21acbf3aea035bbac2c6f877154>

Web of Science Incites Journal Citation Reports Essential Science Indicators EndNote Publius Kopenio

Sign In Help English

Clarivate Analytics

Search Tools Searches and alerts Search History Marked List

Results: 45 (from Web of Science Core Collection)

You searched for: TITLE: (abstract reasoning) ... More

Create Alert

Refine Results

Search within results for...

Filter results by:

☐ Open Access (6)

Refine

Pagination Years

☐ 2019 (2)
☐ 2018 (3)
☐ 2017 (3)
☐ 2016 (4)
☐ 2015 (4)

more options / values... Refine

Web of Science Categories

☐ COMPUTER SCIENCE: ARTIFICIAL INTELLIGENCE (46)

Sort by: Date Times Cited Use Count Relevance More

Select Page Export... Add to Marked List

1. **An abstract, argumentation-theoretic approach to default reasoning**
By Bondarenko, A; Dung, PH; Kowalski, R; et al.
ARTIFICIAL INTELLIGENCE Volume: 93 Issue: 1-2 Pages: 63-101 Published: JUN 1997
Links Free Full Text from Publisher View Abstract

2. **Methods for solving reasoning problems in abstract argumentation - A survey**
By Charwat, Guenther; Dvorak, Wolfgang; Gaggl, Sarah A; et al.
ARTIFICIAL INTELLIGENCE Volume: 220 Pages: 28-63 Published: MAR 2015
Links Free Full Text from Publisher View Abstract

3. **Integration of temporal reasoning and temporal-data maintenance into a reusable database mediator to answer abstract, time-oriented queries: The Tzolkim system**
By Nguyen, JH; Shahar, Y; Tu, SW; et al.
JOURNAL OF INTELLIGENT INFORMATION SYSTEMS Volume: 13 Issue: 1-2 Pages: 121-145 Published: JUL 1999
Links Full Text from Publisher View Abstract

4. **Abstract reasoning for planning and coordination**
By Clement, Bradley J; Durfee, Edmund H; Barrett, Anthony C.
JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH Volume: 28 Pages: 453-515 Published: 2007
Links Free Full Text from Publisher View Abstract

5. **A methodology for designing systems to reason with legal cases using Abstract Dialectical Frameworks**
By Al-Abdulkarim, Latifa; Atkinson, Katie; Bench-Capon, Trevor
ARTIFICIAL INTELLIGENCE AND LAW Volume: 24 Issue: 1 Pages: 1-49 Published: MAR 2016

Analyze Results Create Citation Report

Times Cited: 334 (from Web of Science Core Collection) Usage Count

Times Cited: 43 (from Web of Science Core Collection) Usage Count

Times Cited: 26 (from Web of Science Core Collection) Usage Count

Times Cited: 19 (from Web of Science Core Collection) Usage Count

Times Cited: 17 (from Web of Science Core Collection) Usage Count

Scopus

[Search](#)
[Sources](#)
[Alerts](#)
[Lists](#)
[Help](#)
[SciVal](#)
[Andreas Holzinger](#)

4 document results

[View secondary documents](#)
[View 1 patent result](#)
[View 145987 Mendely Data](#)

TITLE-ABS KEY (meta learning AND memory-augmented AND neural AND networks)

[Edit](#)
[Save](#)
[Set alert](#)
[Set feed](#)

Refine results

[Limit to](#)
[Exclude](#)

Year

☐ 2019 (2)
 ☐ 2018 (1)
 ☐ 2016 (1)

Author name

☐ Bartunov, S. (1)
 ☐ Botvinick, M. (1)
 ☐ Das, A.K. (1)

Analyze search results

[Show all abstracts](#)
[Sort on: Cited by \(highest\)](#)

☐ All
 ☐ Save to Mendeley
 [Download](#)
[View citation overview](#)
[View cited by](#)
[Save to list](#)

	Document title	Authors	Year	Source	Cited by
1	Meta-Learning with Memory-Augmented Neural Networks	Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.	2016	33rd International Conference on Machine Learning, ICM, 2016 4, pp.2740-2751	75
View abstract Related documents					
2	Labeled memory networks for online model adaptation	Shankar, S., Sarawagi, S.	2018	32nd AAAI Conference on Artificial Intelligence, AAAI 2018 pp.4034-4041	3
View abstract Related documents					
3	Text normalization using memory augmented neural networks	Pramanik, S., Hussain, A.	2019	Speech Communication 109, pp.15-23	1
View abstract View at Publisher Related documents					

human-centered.ai (Holzinger Group)

33

Student Research Seminar 2019/2020

Analyse **Sanity Preserver**

Suite in spare time by @kshapathy to accelerate research.

Serving last 86743 papers from cs.CV[CLIGAI]Ne[ne]stat.ML

[User:](#) [Pass:](#) [Login or Create](#)

[New to arxiv-sanity? Check out the introduction video.](#)

Q |

most recent
top recent
top type
friends
discussions
recommended
library


Only show v1

Showing most recent Arxiv papers:

Context-Aware Saliency Detection for Image Retargeting Using Convolutional Neural Networks

Mohd Ahmad, Naoor Karim, Enhorah Esmail
10/17/2019 cs.CV | eess.IV

20 pages, 19 figures



1910.0807v1.pdf

[ehow.csimilar](#) | [discuss](#)






Image retargeting is the task of making images capable of being displayed on screens with different sizes. This work should be done so that high-level visual information and low-level features such as texture remain as intact as possible to the human visual system, while the output image may have different dimensions. Thus, simple methods such as scaling and cropping are not adequate for this purpose. In recent years, researchers have tried to improve the existing retargeting methods and introduce new ones. However, a specific method cannot be utilized to retarget all types of images. In other words, different images require different retargeting methods. Image retargeting has a close relationship to image saliency detection, which is relatively a new image processing task. Earlier saliency detection methods were based on local and global but low-level image information. These methods are called bottom-up methods. On the other hand, newer approaches are top-down and mixed methods that consider the high level and semantic information of the image too. In this paper, we introduce the proposed methods in both saliency detection and retargeting. For the saliency detection, the use of image context and semantic segmentation are examined, and a novel mixed bottom-up, and top-down saliency detection method is introduced. After saliency detection, a modified version of an existing retargeting method is utilized for retargeting the images. The results suggest that the proposed image retargeting pipeline has excellent performance compared to other tested methods. Also, the subjective evaluations on the Pascal dataset can be used as

human-centered.ai (Holzinger Group)

35

Student Research Seminar 2019/2020


 Cornell University
 We gratefully acknowledge support from the Simons Foundation and member institutions.


 arXiv

[Help](#) | [Advanced Search](#)

[Login](#)

Showing 1-1 of 1 results for title: explainable ai medical

[Search v0.5 released 2018-12-20](#)
[Feedback?](#)

☒ Show abstracts
 ☐ Hide abstracts

results per page.
 Sort results by:

1. arXiv:1712.09923 [pdf, other] [\[CS.AI\]](#) [\[stat.ML\]](#)

What do we need to build explainable AI systems for the medical domain?
 Authors: Andreas Holzinger, Chris Biemann, Constantin S. Pattichis, Douglas B. Kell
 Abstract: Artificial intelligence (AI) generally and machine learning (ML) specifically demonstrate impressive practical success in many different application domains, e.g. in autonomous driving, speech recognition, or recommender systems. Deep learning approaches, trained on extremely large data sets or using reinforcement learning have even exceeded human performance in visual tasks, particularly... [View More](#)
 Submitted 28 December, 2017; originally announced December 2017.
 Comments: This is a survey article and section 3.1, draws heavily from arXiv:1706.07979

human-centered.ai (Holzinger Group)

34

Student Research Seminar 2019/2020

YouTube^{AT}

Search _____

We gratefully acknowledge support from
the Open Foundation
and member institutions.

Open access to 1,130,861 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics.
Subject search and browse: [Physics](#) Search Sort Interface Feedback

25 Jan 2016 - A project update, including a brief summary of activities in 2015, has been posted
1 Jun 2016 - New members join arXiv Scientific Advisory Board
See what's "New" pages. Read robots beware before attempting any automated download

Physics

- Astrophysics ([astro-ph new, recent, find](#))
 - Included: Atmospheric and Nonrelativistic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics
- Condensed Matter ([cond-mat new, recent, find](#))
 - Includes: Disordered Systems and Neural Networks; Materials Science: Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter;
- General Relativity and Quantum Cosmology ([gr-qc new, recent, find](#))
 - High Energy Physics - Cosmology and Experiment ([hep-th cos new, recent, find](#))
 - High Energy Physics - Lattice ([hep-lat new, recent, find](#))
 - High Energy Physics - Phénoménologie ([hep-ph new, recent, find](#))
 - High Energy Physics - Theory ([hep-th new, recent, find](#))
- Mathematical Physics ([math-ph new, recent, find](#))
- Numerical Analysis ([num-an new, recent, find](#))
- Nonlinear Sciences ([nlin new, recent, find](#))
 - Biology: Adaptation and Self-Organizing Systems; Cellular Automata and Lattices Games; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
 - Inclusion: Experimental ([nlin ex new, recent, find](#))
 - Nuclear Theory ([nucl-th new, recent, find](#))
- optics ([optics new, recent, find](#))

[arXiv preprint](#) | [help](#) | [about](#) | [contact us](#) | [faq](#) | [privacy policy](#) | [terms & conditions](#) | [donate](#) | [subscribe](#) | [report problem](#)

Introducing arxiv-sanity

51,133 posts • 23 Mar 2016

1.2K 1 SHARE SAVE ...

Andrej Karpathy
24.2K subscribers

SUBSCRIBE

<http://vision.stanford.edu>

human-centered.ai (Holzinger Group)

36

Student Research Seminar 2019/2020

← → ↻ 🏠 https://scholar.google.at/scholar?hl=de&as_sdt=0%2CS&q=explainable+AI&btnG=

Google Scholar explainable AI

Artikel Ungefähr 36 600 Ergebnisse (0,23 Sek.)

Beleibige Zeit
Seit 2019
Seit 2018
Seit 2015
Zeitraum wählen...

Nach Relevanz sortieren
Nach Datum sortieren

Beleibige Sprache
Seiten auf Deutsch

✓ Patente einschließen
✓ Zitate einschließen

Alert erstellen

Explainable artificial intelligence (xai) [PDF] darpa.mil
D Gunning - Defense Advanced Research Projects Agency (DARPA) ... 2017 - darpa.mil
... 2 The Need for **Explainable AI** User - Why did you do that ... **Explainable AI** will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners **AI** System <http://istaverse.com/> ©2007–2017 Istaverse Ltd ...
☆ 99 Zitiert von: 280 Ähnliche Artikel Alle 6 Versionen In EndNote importieren 30

What do we need to build explainable AI systems for the medical domain? [PDF] arxiv.org
A Holzinger, G Bieemann, GS Dettliche - arXiv preprint arXiv ... 2017 - arxiv.org
Artificial intelligence (**AI**) generally and machine learning (**ML**) specifically demonstrate impressive practical success in many different application domains, eg in autonomous driving, speech recognition, or recommender systems. Deep learning approaches, trained ...
☆ 99 Zitiert von: 79 Ähnliche Artikel Alle 4 Versionen In EndNote importieren 30

Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences [PDF] arxiv.org
T Miller, P Eizen, L Sonnenberg - arXiv preprint arXiv 1712.00547, 2017 - arxiv.org
In his seminal book *The Inmates are Running the Asylum*, Why High-Tech Products Drive Us Crazy And How To Restore The Sanity [2004, Sams Indianapolis, IN, USA], Alan Cooper argues that a major reason why software is often poorly designed (from a user perspective) ...
☆ 99 Zitiert von: 69 Ähnliche Artikel Alle 6 Versionen In EndNote importieren 30

What does explainable AI really mean? A new conceptualization of perspectives [PDF] arxiv.org
D Doran, S Schulz, TR Bessag - arXiv preprint arXiv 1710.00794, 2017 - arxiv.org
We characterize three notions of **explainable AI** that cut across research fields: opaque systems that offer no insight into its algorithmic mechanisms; interpretable systems where users can mathematically analyze its algorithmic mechanisms; and comprehensible ...
☆ 99 Zitiert von: 46 Ähnliche Artikel Alle 13 Versionen In EndNote importieren 30

Explainable AI: the new 42? [PDF] inria.fr
R Goebel, A Chander, K Holzinger, E Lecue - ... Domain Conference for ... 2018 - Springer
Explainable AI is not a new field. Since at least the early exploitation of CS Pierce's abductive reasoning in expert systems of the 1980s, there were reasoning architectures to support an explanation function for complex **AI** systems, including applications in medical ...
☆ 99 Zitiert von: 25 Ähnliche Artikel Alle 15 Versionen In EndNote importieren 30

THOMSON REUTERS
ENDNOTEL^AT_EXBIB_TE_X

06 How to manage Papers?

EndNote X6 - 3D LIBRARY_6730-16030fah.ent

File Edit References Groups Tools Window Help

Search Options

Search Whole Library Match Case Match Words

Author Contains Pass

And Year Contains

And Title Contains

Label Author Year Title Rating Journal

7639 Holzinger(AI)2... Holzinger, A; Plass, M; Holzinger, K; Crisan, G... 2016 Towards interactive Machine Learning (IML): Applying Ant Colony Alg... Springer Lecture No

EndNote X6 - 3D LIBRARY_6730-16030fah.ent

File Edit References Groups Tools Window Help

Search Options

Search Whole Library Match Case Match Words

Author Contains Pass

And Year Contains

And Title Contains

Label Author Year Title Rating Journal

7639 Holzinger(AI)2... Holzinger, A; Plass, M; Holzinger, K; Crisan, G... 2016 Towards interactive Machine Learning (IML): Applying Ant Colony Alg... Springer Lecture No

Reference Preview Attached PDFs

Holzinger, A, Plass, M, Holzinger, K, Crisan, G, Pintea, C & Palade, V. 2016. Towards interactive Machine Learning (IML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach. Springer Lecture Notes in Computer Science LNCS 9817. Heidelberg, Berlin, New York: Springer, pp. 81-95, doi:10.1007/978-3-319-45507-56.

Reference Preview Attached PDFs

[@incollection]Holzinger(AI)2016-MLEExperiment,
year = (2016),
author = (Holzinger, A and Plass, M and Holzinger, K and Crisan, G and Pintea, CM and Palade, V.),
title = (Towards interactive Machine Learning (IML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach),
booktitle = (Springer Lecture Notes in Computer Science LNCS 9817),
publisher = (Springer),
address = (Heidelberg, Berlin, New York),
pages = (81-95),
abstract = (Most Machine Learning (ML) researchers focus on automatic Machine Learning (ML) where great advances have been made. However, sometimes, for example in health informatics, we are confronted with a small number of data sets or rare events, and with complex problems where ML approaches fail or deliver unsatisfactory results. Here, interactive Machine Learning (IML) may be of help and the "human-in-the-loop" approach may be beneficial in solving computationally hard problems, where human expertise can help to reduce an exponential search space through heuristics. In this paper, experiments are discussed which help to evaluate the effectiveness of the ML "human-in-the-loop" approach, particularly in opening the "black box", thereby enabling a human to directly and indirectly manipulate and interacting with an algorithm. For this purpose, we selected the Ant Colony Optimization (ACO) framework and use it on the Traveling Salesman Problem (TSP) which is of high importance in solving many practical problems in health informatics, e.g. in the study of proteins.),
doi = (10.1007/978-3-319-45507-56).

Holzinger, A., Plass, M., Holzinger, K., Crisan, G., Pintea, C. & Palade, V. 2016. Towards interactive Machine Learning (IML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach. Springer Lecture Notes in Computer Science LNCS 9817. Heidelberg, Berlin, New York: Springer, pp. 81-95, doi:10.1007/978-3-319-45507-56.

07 Publication targets



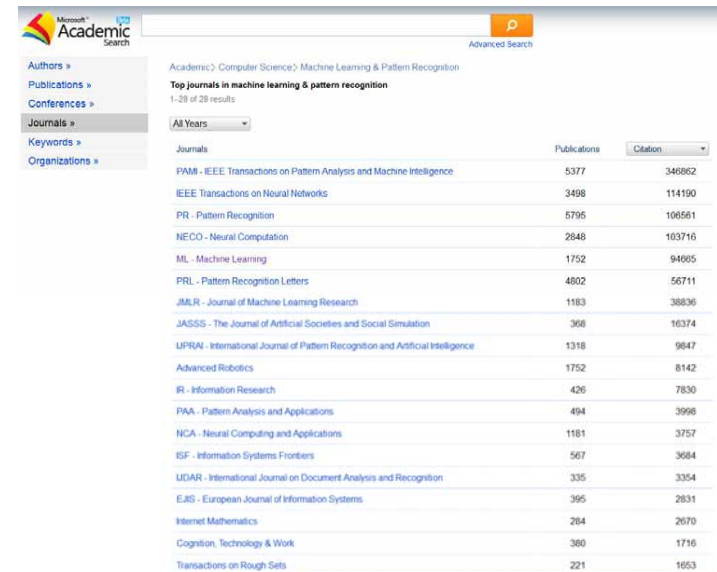
<https://academic.microsoft.com/conferences/41008148>



human-centered.ai (Holzinger Group)

45

Student Research Seminar 2019/2020



<http://academic.research.microsoft.com/RankList?entitytype=4&topdomainid=2&subdomainid=6&last=0&orderby=1>

human-centered.ai (Holzinger Group)

46

Student Research Seminar 2019/2020

08 How to write a Paper?

- Stop Reading! Start Writing!
- Read along when writing!
- SQ3R-Method:
 - Survey (read title, abstract, conclusion, subheadings)
 - Question (what are the major insights of this paper?)
 - Read: with regard to the question above
 - Recite: summarize with your own words
 - Review: Try to reflect the major insights of the paper
- Do not waste time! Be economic! It is simply impossible to read everything and all!

- 1. Set goal (e.g. to bring paper into conference x or journal y) – write a preliminary (!) title and abstract
- 2. Study published work related to your topic
- 3. A good start is on the “future outlook” sections of published papers – outline intended work on one single page (birds eye view – eagle top view)
- 4. Start Writing! Discuss the related work and the theoretical background – leave gaps
- 5. Now bring in your work, experiments and results
- 6. Write Introduction, Conclusion, revise abstract, revise the title accordingly
- 7. Submit your paper
- 8. Carefully read the reviews, revise accordingly

- What is the problem? Is it challenging?
- How can the problem be solved – alternative methods, background, related work?
- How well is the problem solved (evaluation)?
- How useful is the result to the intended readers?
- Example:
 - We propose a method ... this is important because ... we solve this problem via ... finally we demonstrate that our method outperforms the state-of-the-art ...

On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation

By: Bach, S. (Bach, Sebastian)^{1,2,†}; Binder, A. (Binder, Alexander)^{2,5,†}; Montavon, G. (Montavon, Gregoire)^{2,†}; Klauschen, F. (Klauschen, Frederick)^{2,†}; Müller, K.R. (Müller, Klaus-Robert)^{2,4,†}; Samek, W. (Samek, Wojciech)^{1,2,†}

View Web of Science ResearcherID and ORCID

PLOS ONE
Volume: 10 Issue: 7
Article Number: e0130140
DOI: 10.1371/journal.pone.0130140
Published: JUL 10 2015
Document Type: Article
View Journal Impact

Abstract

Understanding and interpreting classification decisions of automated image classification systems is of high value in many applications, as it allows to verify the reasoning of the system and provides additional information to the human expert. Although machine learning methods are solving very successfully a plethora of tasks, they have in most cases the disadvantage of acting as a black box, not providing any information about what made them arrive at a particular decision. This work proposes a general solution to the problem of understanding classification decisions by pixel-wise decomposition of nonlinear classifiers. We introduce a methodology that allows to visualize the contributions of single pixels to predictions for kernel-based classifiers over Bag of Words features and for multilayered neural networks. These pixel contributions can be visualized as heatmaps and are provided to a human expert who can intuitively not only verify the validity of the classification decision, but also focus further analysis on regions of potential interest. We evaluate our method for classifiers trained on PASCAL VOC 2009 images, synthetic image data containing geometric shapes, the MNIST handwritten digits data set and for the pre-trained ImageNet model available as part of the Caffe open source package.

Keywords
KeyWords Plus: NEURAL-NETWORKS; IMAGE; FEATURES; MODEL

Author Information
Reprint Address: Bach, S (reprint author)
† Fraunhofer Heinrich Hertz Inst, Machine Learning Grp, Berlin, Germany.

Citation Network
In Web of Science Core Collection
170
Times Cited
Create Citation Alert

All Times Cited Counts
171 in All Databases
See more counts

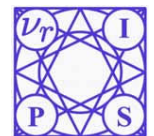
58
Cited References
View Related Records

Most recently cited by:
Zhang, JiaWei; Zhao, XiaoJian; Zhang, Jiyi; et al.
Agency MIS Prepayment Model Using Neural Networks.
JOURNAL OF STRUCTURED FINANCE (2019)
Yuan, Man; Liu, Zhi; Wang, Fan; et al.
Rethinking Labelling in Road Segmentation.
INTERNATIONAL JOURNAL OF REMOTE SENSING (2019)

- The following example is in accordance with the NIPS*) committee, credit to: Andrew Ng, Peter Dayan, Daphne Koller, Sebastian Thrun, Bruno Olshausen, Yair Weiss, Bernhard Schölkopf, Max Welling & Zoubin Ghahramani
- <https://nips.cc/Conferences/2016/PaperInformation/EvaluationCriteria>

International Conference on Neural Information Processing Systems, founded in 1987 as NIPS, and renamed 2018 to NeurIPS

The NeurIPS Conference is organized by the NeurIPS Foundation, established by Ed Posner (1933-1993) in 1987, and chaired by Terrence Sejnowski (1947-) since 1993. The board of trustees consists of previous general chairs of the NeurIPS Conference.



The papers are collected here: <https://papers.nips.cc>

Very interesting and where you can learn a lot is the open review system, see next slide ...

Remember: The best way of reading a paper is to review a paper!

A simple neural network module for relational reasoning

Adam Santoro*
adasantoro@google.com

David Raposo*
draposo@google.com

David G.T. Barrett
barrettdavid@google.com

Mateusz Malinowski
mateusz@google.com

Razvan Pascanu
razp@google.com

Peter Battaglia
peterbattaglia@google.com

Timothy Lillicrap
lillicrap@google.com

London, United Kingdom
countzero@google.com

Abstract

Relational reasoning is a central component of generally intelligent behavior, but has proven difficult for neural networks to learn. In this paper we describe how to use Relation Networks (RNs) as a simple plug-and-play module to solve problems that fundamentally hinge on relational reasoning. We tested RN-augmented networks on three tasks: visual question answering using a challenging dataset called CLEVR, on which we achieve state-of-the-art, super-human performance; text-based question answering using the bAbI suite of tasks; and complex reasoning about dynamical physical systems. Then, using a curated dataset called Sort-of-CLEVR we show that powerful convolutional networks do not have a general capacity to solve relational questions, but can gain this capacity when augmented with RNs. Thus, by simply augmenting convolutions, LSTMs, and MLPs with RNs, we can remove computational burden from network components that are not well-suited to handle relational reasoning, reduce overall network complexity, and gain a general ability to reason about the relations between entities and their properties.

Authors

- Adam Santoro
- David Raposo
- David G. Barrett
- Mateusz Malinowski
- Razvan Pascanu
- Peter Battaglia
- Timothy Lillicrap

Conference Event Type: Poster

Abstract

Relational reasoning is a central component of generally intelligent behavior, but has proven difficult for neural networks to learn. In this paper we describe how to use Relation Networks (RNs) as a simple plug-and-play module to solve problems that fundamentally hinge on relational reasoning. We tested RN-augmented networks on three tasks: visual question answering using a challenging dataset called CLEVR, on which we achieve state-of-the-art, super-human performance; text-based question answering using the bAbI suite of tasks; and complex reasoning about dynamical physical systems. Then, using a curated dataset called Sort-of-CLEVR we show that powerful convolutional networks do not have a general capacity to solve relational questions, but can gain this capacity when augmented with RNs. Thus, by simply augmenting convolutions, LSTMs, and MLPs with RNs, we can remove computational burden from network components that are not well-suited to handle relational reasoning, reduce overall network complexity, and gain a general ability to reason about the relations between entities and their properties.

A simple neural network module for relational reasoning

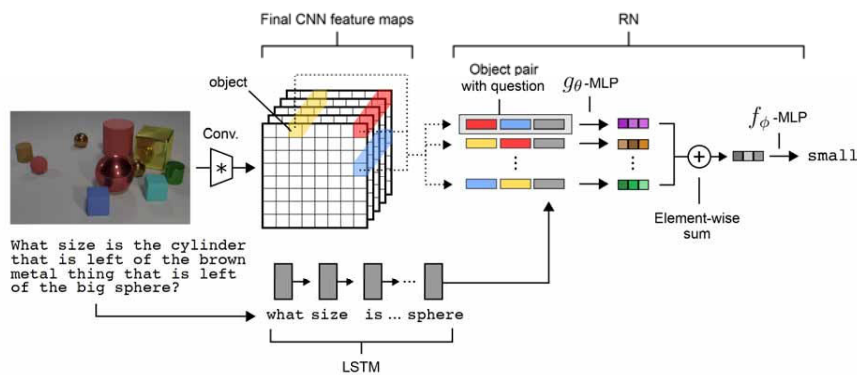
A. Santoro, D. Raposo, D.G. Barrett, et al. - Advances in neural information processing systems 30 (NIPS 2017) - papers.nips.cc

The design philosophy behind RNs is to constrain the functional form of a neural network so that it captures the core common properties of relational reasoning - built-in to CNNs, and the capacity to reason about sequential dependencies is built into recurrent neural networks

☆ 99 Zitiert von: 430 Ähnliche Artikel Alle 6 Versionen In EndNote importieren 90

Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P. & Lillicrap, T. A simple neural network module for relational reasoning. In: Guyon, Isabelle, Luxburg, Ulrike V., Bengio, Samy, Wallach, Hannah, Fergus, Rob, Vishwanathan, Svn & Garnett, Roman, eds. Advances in neural information processing systems (NIPS), (2017) Long Beach (CA). Neural Information Processing Society.

<http://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning>



Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia & Timothy Lillicrap. A simple neural network module for relational reasoning. In: Guyon, Isabelle, Luxburg, Ulrike V., Bengio, Samy, Wallach, Hannah, Fergus, Rob, Vishwanathan, Svn & Garnett, Roman, eds. Advances in neural information processing systems (NIPS), 2017 Long Beach (CA). Neural Information Processing Society, 4967-4976.

A simple neural network module for relational reasoning

In Artikeln mit Zitaten suchen

Graph attention networks

P. Veličković, G. Cucurull, A. Casanova, et al. - arXiv preprint arXiv:1710.10903, 2017 - arxiv.org

We present graph attention networks (GATs), novel neural network architectures that operate on graph-structured data, leveraging masked self-attentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations. By ...

☆ 99 Zitiert von: 550 Ähnliche Artikel Alle 7 Versionen In EndNote importieren 90

Non-local neural networks

X. Wang, B. Ghanem, A. Gupta, et al. - Proceedings of the IEEE, 2018 - openaccess.thecvf.com

Both convolutional and recurrent operations are building blocks that process one local neighborhood at a time. In this paper, we present non-local operations as a generic family of building blocks for capturing long-range dependencies. Inspired by the classical non-local ...

☆ 99 Zitiert von: 539 Ähnliche Artikel Alle 7 Versionen In EndNote importieren 90

Deep reinforcement learning: An overview

Y. Li - arXiv preprint arXiv:1701.07274, 2017 - arxiv.org

We give an overview of recent exciting achievements of deep reinforcement learning (RL). We discuss six core elements, six important mechanisms, and twelve applications. We start with background of machine learning, deep learning and reinforcement learning. Next we ...

☆ 99 Zitiert von: 266 Ähnliche Artikel Alle 4 Versionen In EndNote importieren 90

Learning to compare: Relation network for few-shot learning

F. Sung, Y. Yang, L. Zhang, T. Xiang, et al. - Proceedings of the IEEE, 2018 - openaccess.thecvf.com

We present a conceptually simple, flexible, and general framework for few-shot learning, where a classifier must learn to recognise new classes given only few examples from each. Our method, called the Relation Network (RN), is trained and tested end-to-end from scratch. During ...

☆ 99 Zitiert von: 233 Ähnliche Artikel Alle 12 Versionen In EndNote importieren 90

Film: Visual reasoning with a general conditioning layer

E. Pérez, E. Strub, H. De Vries, Y. Dumoulin, et al. - Thirty-Second AAAI, 2018 - aaai.org

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple, feature-wise affine transformation based on conditioning information. We show that ...

☆ 99 Zitiert von: 151 Ähnliche Artikel Alle 6 Versionen In EndNote importieren 90

- The paper proposes a plug and play module (called Relation Networks (RNs)) specialized for relational reasoning. The module is composed of Multi Layer Perceptrons and considers relations between all pairs of objects. The proposed module when plugged into traditional networks achieves state of the art performance on the CLEVR visual question answering dataset, state of the art (with joint training for all tasks) on the bAbI textual question answering dataset and high performance (93% on one task and 95% on another) on a newly collected dataset of simulated physical mass-spring systems. The paper also collects a dataset similar to CLEVR to demonstrate the effectiveness of the proposed RNs for relational questions.
- Strengths:**
 - 1. The proposed Relation Network is a novel neural network specialized for relational reasoning. The success of the proposed network is extensively shown by experimenting with three different tasks and clearly analyzing the effectiveness for relational questions by collecting a novel dataset similar to CLEVR.
 - 2. The proposed RNs have been shown to be able to work with different forms of input – explicit state representations as well as features from a CNN or LSTM.
 - 3. The paper is well written and the details of model architecture including hyperparameters are provided.
 - 4. As argued in the paper, I agree that relational reasoning is central to intelligence and since RNs are shown to be able to achieve this reasoning and a result perform better at tasks requiring such reasoning than existing networks, RNs seem to be of significant importance for designing reasoning networks.
- Weaknesses:**
 - 1. Could authors please analyze and comment on how complicated relations can be handled by RNs. Is it the case that RNs perform well for single hop relations such as "what is the color of the object closest to the blue object" which requires reasoning about only one hop relation (distance between blue object and all other objects), but not so well for multiple hop relations such as "What shape is the small object that is in front of the yellow matte thing and behind the gray sphere?". From the failure cases in table 1 of supplementary material, it seems that the model has difficulty in answering questions involving multiple hops of relations.
 - 2. L203-204, it is not clear to me what do authors mean by "we tagged ... support set". Is this referring to some form of human annotation? If so, could authors please elaborate on what happens at test time?
 - 3. All the datasets experimented with in the paper are synthetic datasets. Could authors please comment on how they expect the RNs to work on real datasets such as the VQA dataset from Antol et al.?
- Post-rebuttal comments:
 - Authors have provided satisfactory response to my question about multi-hop reasoning. However, I would still like to see experiments on real VQA dataset to see how effective RNs are at dealing with the amount of variation real datapoints show (in vision as well as in language). So it would be great if authors could include results on the VQA dataset (Antol et al., ICCV 2015) in camera-ready.

http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips30/reviews/2565.html

human-centered.ai (Holzinger Group)

57

Student Research Seminar 2019/2020

- Reviewer 2
- This paper presents the relational network module, which when included as part of a larger network architecture is able to essentially solve the CLEVR VQA task despite its simplicity. The model is also tested over other synthetic tasks involving both vision and language; I hope that in the future the authors can also demonstrate its effectiveness on real-world tasks. While at this point I don't think results on bAbI are particularly informative, especially with the strange way the bAbI task is set up in this paper (what does "up to 20 support sentences" mean? was there some rule-based or ML method used to select these support sentences? why not use all sentences as support?), the model is an interesting and effective way to think about relational QA problems, and I hope that the paper is accepted. That said, I have some questions/comments about the paper:
 - were functions other than simple summation in Eq.1 experimented with?
 - what is the function of the "arbitrary coordinate" in the "dealing with pixels" section? How is it "arbitrary" if it is supposed to indicate relative position? there is not enough detail to understand how it is implemented, and Figure 2 offers no insight despite being referenced. Is this crucial to making the RN work? If so, it needs to be stated in the paper.
 - is the reason why all support sentences aren't used in bAbI due to computational constraints (since some of the bAbI tasks have large contexts), and is this a potential limitation of the RN?
 - the model doesn't make as much sense for NLP tasks where "objects" aren't clearly defined, and there is an obvious order to the objects (there is discourse-level information in the sequential ordering of sentences, even in bAbI). Couldn't the model be applied across different units of text (e.g., treat words, phrases, sentences, paragraphs all as objects)? Any thoughts on how this could be implemented in for example bAbI?
 - the CNN used in the CLEVR experiments is very simple compared to prior work, which utilized VGG features. since the CLEVR images are artificially constructed, it strikes me that a simpler CNN is better suited for the problem. What happens if you run the RN on CLEVR but utilize feature maps from the last conv layer of VGG as objects? this would be a more fair comparison than what is described in the paper.

http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips30/reviews/2565.html

human-centered.ai (Holzinger Group)

58

Student Research Seminar 2019/2020

OpenReview.net

Search ICML 2017 WHI

Login

Go to ICML 2017 WHI homepage

Interpretable Active Learning

Richard L. Phillips, Kyu Hyun Chang, Sorelle Friedler

17 Jun 2017 (modified: 19 Jun 2017) ICML 2017 WHI Submission Readers: Everyone

Abstract: Active learning has long been a topic of study in machine learning. However, as increasingly complex and opaque models have become standard practice, the process of active learning, too, has become more opaque. There has been little investigation into interpreting what specific trends and patterns an active learning strategy may be exploring. This work expands on the Local Interpretable Model-agnostic Explanations framework (LIME) to provide explanations for active learning recommendations. We demonstrate how LIME can be used to generate locally faithful explanations for an active learning strategy, and how these explanations can be used to understand how different models and datasets explore a problem space over time. We propose a measure for uncertainty bias based on disparate impact that allows further exploration of the relative exploitation of different data subgroups. We combine the LIME framework with the uncertainty bias metric to demonstrate how clusters of unlabeled points can be made automatically based on common sources of uncertainty. We show that this allows for an interpretable explanation of what an active learning algorithm is learning as points with similar sources of uncertainty have their uncertainty bias resolved.

TLDR: Attempt to analyze trends in the concepts active learning is exploring

Keywords: Active learning, interpretable machine learning

8 Replies

Show

all

from

everybody

Interesting paper

ICML 2017 WHI Paper14 AnswerReview3

29 Jun 2017 (modified: 05 Jul 2017) ICML 2017 WHI Paper14 Official Review Readers: Everyone

Rating: 5: Top 50% of accepted papers, clear accept

Review: This work is highly significant as most active learning methods are opaque. The authors provides quantification of the uncertainty bias. The paper experimentally demonstrate how to create and track uncertain groups and fetch labels.

Confidence: 2: The reviewer is fairly confident that the evaluation is correct

Re: Interesting paper

Richard L. Phillips

30 Jun 2017 (modified: 30 Jun 2017) ICML 2017 WHI Comment Readers: Everyone

Comment: This is a correct summary. Thank you for reviewing our paper.

Review of "Interpretable active learning". Interesting idea, but execution needs improvement.

ICML 2017 WHI Paper14 AnswerReview2

28 Jun 2017 (modified: 05 Jul 2017) ICML 2017 WHI Paper14 Official Review Readers: Everyone

<https://openreview.net/forum?id=H1jpN5GmZ¬elid=H1SbKvamW>

human-centered.ai (Holzinger Group)

59

Student Research Seminar 2019/2020

Review of "Interpretable active learning". Interesting idea, but execution needs improvement.

ICML 2017 WHI Paper14 AnswerReview2

28 Jun 2017 (modified: 05 Jul 2017) ICML 2017 WHI Paper14 Official Review Readers: Everyone

Rating: 4: Marginally above acceptance threshold

Review: Summary of the paper:
The paper proposes to adapt the LIME local interpretation approach to the active learning problem. Specifically, the authors consider "uncertainty sampling" version of active learning, and apply the regression version of LIME to provide local explanations for uncertainty of data points. The authors also propose to track "uncertainty bias" in specially defined regions of the feature space.

Overall assessment: The proposed idea is interesting – both using LIME to explain active learning, and to track regions of the feature space with unusual uncertainties. However, the description of the approach is done very casually, and after re-reading it twice I still have doubts whether I understood exactly what is proposed. I would encourage the authors to make the discussion much clearer and more precise.

Confusing terms:

Sec 4: "Combinations of LIME explanations of uncertainty" are uncertainty regions. What is a combination here? Is it a linear / convex combination, or a set of points with their explanations? If it's a set – are these mutually exclusive or overlapping? This is very loosely defined here.

Sec 5.5: This section is impenetrable! You need to rewrite it much more clearly, and define exactly what you propose, and your assumptions. Do you assume that the entire data set (labeled and unlabeled data points) is available at the start? What specifically do you cluster – labeled / unlabeled / full data? "Use label as its own dimension"???? What does it mean? Do you mean that after you use k-means – you use the cluster label as an interpretable feature to feed to LIME? What is a "value of a point"? What is a "constant"? It's the first time this term is used. So far I've been able to guess what you're trying to do – but in the remaining part I get entirely lost. How do you choose the number of clusters. What are top-uncertainty constraints? Please rewrite.

e) If you use k-means to split the data set into regions – how often are these regions interpretable? E.g. can you identify the region with some values of "protected demographic"? Or is just some fairly arbitrary set of data points? My concern: k-means treats all features as equal in weight, and when there are many features it won't have any interpretation.

f) When you apply k-means to cluster features – e.g. a level distribution of some continuous features – do you assume in your experiments that nearby "buckets" are "closer" than further buckets? E.g. in your definition, is the distance from 20-year-olds to 30-year-olds bucket, the same as from 30-year-olds to 40-year-olds, or is it closer? Doesn't it make more sense to define specific groups you're concerned with (e.g. aligned with gender, economic status) – and track uncertainty bias there directly – maybe in addition to k-means? Or at least use subspace clustering which is more "aligned" with individual features.

g) the description of the chemical data set and the difficulty you ran into is not clear. It sounds like with more features it's harder to gain insight?

h) Why do you use ridge regression? It's a poor way to select top features. Lasso / group lasso / stepwise regression n.c.c. would be a much better choice.

Figure 1: It'd be nice to label the axes on the plot, and to show in the figure the regions defined in text on the right of the figure.

Pro-Public example: "Each data point was given two uncertainty labels" What are the two labels? Do you mean that it was given one of the two uncertainty labels (1 or 3)? Otherwise I have no idea.

"On the top most explored clusters uncertainty bias tended downwards." I don't see what the opposite – where uncertainty bias tends downward for most "unexplored clusters"

"Uncertainty labels for points that are frequently queried truly correlate with the resolved uncertainty." 177 is an active learning setting – don't you query a point just once? Is this the multiple repeated experiments? How are these conducted? The set up needs to be described fully.

Confidence: 2: The reviewer is fairly confident that the evaluation is correct

My understanding: You first partition the labeled data set (or entire data set) into some number of regions using k-means, and bias each data point in the cluster to its cluster centroid. (Note that the k-means cluster center may not be a valid data point (e.g. for text / images) – you may need to use ensemble clustering instead – see ref 15). Then you use cluster labels as features to run ridge regression to predict the uncertainty for each unlabeled data point. These are the "interpretable" features in the LIME framework – which differ from the original features from the data set. Is that roughly correct?

Specific comments:

a) There are many other widely used choices for active learning – beyond uncertainty sampling. For example a combination of uncertainty and diversity, or coverage. In many classification problems uncertainty sampling would select samples from extremely rare regions – that do not materially improve the rate of sample accuracy "in the wild". So one needs to select fairly "typical examples" which still are uncertain. Does your approach extend to that setting? How?

b) One can incorporate costs of queries (e.g. chemical compounds already purchased) directly – without interpretability. Using cost-weighted active learning, how does interpretability help there? c.f. see ref. **Klein, Gary M., and Foster Provost. "Learning when training data are costly: The effect of class distribution on bias induction." *Journal of Artificial Intelligence Research* 19 (2003): 315-364.

c) What is the benefit of defining the uncertainty bias by quantifying uncertainty into two values? Do you gain anything? Can't you just define it the same exact way, but using raw uncertainty values? e.g. 1 - average (L1) / average (L2, RBF) etc.

d) There is some work on using active learning for chemical reactions, and they also mention in passing interpretability (briefly at the end). **Wang, Matthew X., et al. "Active learning in the drug discovery process." *NIPS*. 2002.

e) I have ref for example clustering Large-scale Submodular Greedy Maximization Selection with Structured Similarity Matrices, L. Sun et al. IJML 2016.

<https://openreview.net/forum?id=H1jpN5GmZ¬elid=H1SbKvamW>
human-centered.ai (Holzinger Group)

60

Student Research Seminar 2019/2020

Interpretable Active Learning

Richard L. Phillips¹ Kyu Hyun Chang² Sorelle Friedler¹

Abstract

Active learning has long been a topic of study in machine learning. However, as increasingly complex and opaque models have become standard practice, the process of active learning, too, has become more opaque. There has been little investigation into interpreting what specific trends and patterns an active learning strategy may be exploiting. This work expands on the Local Interpretable Model-agnostic Explanations framework (LIME) to provide explanations for active learning recommendations. We demonstrate how LIME can be used to generate locally faithful explanations for an active learning strategy, and how these explanations can be used to understand how different models and datasets explore a problem space over time. We propose a measure for uncertainty bias based on disparate impact that allows further exploration of the relative exploitation of different data subgroups. We combine the LIME framework with the uncertainty bias metric to demonstrate how clusters of unlabeled points can be used to automatically based on common sources of uncertainty. We show that this allows for an interpretable explanation of what an active learning algorithm is learning as points with similar sources of uncertainty have their uncertainty bias reduced.

1. Introduction

The importance of interpretability and explainability of machine learned decisions has recently been an area of active interest, with the EU even declaring what has been called a "right to an explanation" (Goodman

& Flaxman, 2016). Recent work on interpretability has included both local explanations about an individual's decision (Bibetto et al., 2016) and global explanations about the model's actions overall, and has included interpretable techniques in clustering (Chen et al., 2016), integer programming (Zeng et al., 2016), rule sets (Wang & Bollen, 2015), and methods for understanding deep nets (Zeiler & Fergus, 2014; Le et al., 2011) in addition to historical work on decision trees (Quinlan, 1993) and random forests (Breiman, 2001). In these traditional machine learning contexts, the focus of interpretability has been twofold: first on the receiver of the decision ("why was I rejected for this job?") and second on the model creator ("why is my model giving these answers?").

Here, we extend this interest in interpretability to active learning, a domain in which the explanation is additionally of interest to the learner ("why am I being asked these questions and why is it worth it to answer?"). Since active learning is generally applied in scenarios such as drug discovery where it is expensive (whether in terms of time or money) to label a query, the learner in these contexts is often a domain expert in their own right (e.g., a chemist). Given this, we ask these questions and why is it worth it to answer?"). Since active learning is generally applied in scenarios such as drug discovery where it is expensive (whether in terms of time or money) to label a query, the learner in these contexts is often a domain expert in their own right (e.g., a chemist). Given this, we ask these questions and why is it worth it to answer?"). Since active learning is generally applied in scenarios such as drug discovery where it is expensive (whether in terms of time or money) to label a query, the learner in these contexts is often a domain expert in their own right (e.g., a chemist). Given this, we ask these questions and why is it worth it to answer?").

1.1. Results

We demonstrate how active learning choices can be made more interpretable to non-experts. Using per query explanations of uncertainty, we develop a system that allows experts to choose whether to label a query. This allows experts to incorporate domain knowledge and their own interests into the labeling process. For example, in the case of a chemist's knowledge of a chemical system, this might allow a model to focus on the reactions of interest to the chemist, the ones for which reagents are already purchased, or even take advantage of the chemist's existing knowl-

Interpretable Active Learning

edge to learn targeted information faster. Indeed, we demonstrate the potential for such expert-driven active learning systems to outperform traditional active learning strategies.

In addition, we introduce a quantified notion of uncertainty bias, the idea that an algorithm may be less certain about its decisions on some data clusters than others. In the context of decision-making about people, this may mean that some protected groups (e.g., races or genders) may receive less favorable decisions due to risk aversion (Goodman & Flaxman, 2016). In the context of active learning, this means that those groups are more likely to be targeted for exploratory queries in order to improve the model. We combine this idea with the explanations generated per query to describe the groups most targeted by uncertainty bias.

2. Related Work

Active Learning. Active learning has a long history that is detailed in this comprehensive survey (Settle, 2009). Our work will focus on explaining query uncertainty. Uncertainty querying for active learning was first proposed in 1994 by Lewis and Gale (Lewis & Gale, 1994). Since then, it has become perhaps the most common strategy for active learning and several strategies for quantifying uncertainty have been developed (Settle, 2009). Strategies used to quantify uncertainty for actively learning multi-class classification problems include selecting the sample with the minimum maximum-class probability, selecting the sample with the minimum difference in probabilities between the two most probable classes or by choosing the sample with maximal label entropy. All three of the above strategies are equivalent for binary classification tasks, such as the tasks we focus on in this paper (Settle, 2009).

3. Local Interpretable Model-Agnostic Explanations

We will build specifically on a method for creating local explanations introduced in (Ribeiro et al., 2016). Local Interpretable Model-Agnostic Explanations (LIME) is an algorithm for offering prediction explanations for individual predictions. This works well on even very complex models by training an interpretable model on the local space around a prediction. This local approximation is useful for creating associations of factors that are influential in a model's prediction. For a given predicted instance, LIME generates a perturbed sample set in the neighborhood of the in-

stance. Then, based on the sample and the model predictions, LIME searches for the most interpretable model and derives an explanation.

4. Explaining Active Learning Queries

The goal of this work is to explain, beyond the attribute and specific data points queried, a strategy to understand what uncertainty an active learning algorithm is attempting to resolve and to determine whether any subgroups need to be monitored during an active learning run.

Toy example. An example multi-class classification problem is used to explore explanations of uncertainty: Four Gaussian distributions with unit variance are centered at $(-3, -3)$, $(3, -3)$, $(3, 3)$, and $(-3, 3)$. The Gaussians are assigned labels such that the first two represent one class and the second two. Initially, 50 points are randomly selected from the Gaussian at $(-3, -3)$ and $(3, 3)$ to be labeled. The points have been purposefully drawn in such a way so to label some of the points Gaussians centered in the second and fourth quadrants. An initial logistic regression model, W , is trained on the initial 50 labeled points. Based on the resulting model of the probability distribution, the certainty scores across the problem space are mapped. The labeled points, decision boundary, and certainty scores can be seen in Figure 1.

Using LIME (Ribeiro et al., 2016), we can ask for locally-faithful weighted explanations of the certainty values provided by W . We will refer to combinations of LIME explanations of uncertainty as "uncertainty regions." These regions can be useful for grouping points together based on identical sources of uncertainty. In these cases, we would expect points explained in a given uncertainty region to increase the certainty we have about points with the same sources of uncertainty.

5. Identifying Uncertainty Bias

In situations where some instance populations are smaller (minority groups) or where the initial training data distribution is skewed, the active learner may prefer queries that are disproportionately drawn from a single region (or population group). For example, in our toy example above, we see that upper left quadrant is underrepresented in the labeled dataset. The points in this region have higher uncertainty and were more likely to be targeted for active learning queries. In order to understand both what and how an active learning method is learning and whether there is a disparate impact among groups targeted to be labeled, it

Start writing! Reading follows automatically

Concrete Writing (1): A few notes

- 0) **Clear writing** (without any "schnick-schnack"), technical soundness, correctness and the reference of previous work is imperative and shall not be explicitly mentioned!
- 1) Authors are required to submit one or more **keywords**, which are used to assign area chairs and reviewers. However, keywords do not bind any committee in assigning reviewers to the paper. For example, if a paper proposes a new algorithm, but contains no empirical assessment, marking it as a learning theory paper will not necessarily lead to more likely acceptance!
- 2) NIPS is an interdisciplinary conference that covers both natural and synthetic neural information processing systems. It is often the case that strong NIPS papers appeal to both parts of the community: for example, by using modern analysis methods developed on the synthetic side to study natural systems, or by investigating algorithmic aspects of methods used by natural systems. While a broad appeal tends to strengthen a NIPS submission, there are also many strong NIPS papers that are more specialized, and thus only fall into one of the two categories described below.

Concrete Writing on the example of NeurIPS (NIPS)

- Examples of papers: a paper proposing a new learning algorithm; one that describes a solution to a difficult application; or one that proves bounds on the error of some learning method. Such papers are expected to make significant (i) algorithmic, and/or (ii) application, and/or (iii) theoretical contributions.
- NIPS (and any other conference and journal ;-)) seeks to publish papers that **will have a high impact – (measured by "citations")** within the machine learning research community, and beyond. Papers will therefore be evaluated on the basis of the following five criteria:
- 1 **Novelty of algorithm.** For example, a paper that gives an elegant new derivation for an algorithm; or one that proposes a new approach to an existing problem.
- 2 **Novelty of application/problem.** For example, a paper that addresses an important application that has heretofore been little-studied at NIPS and beyond. Or, one that introduces a novel machine learning problem (some past examples include ICA and structured prediction) and proposes an algorithm for it.
- 3 **Difficulty of application.** For example, an application of machine learning to a difficult, important, and "real-world" application, that takes into account the full complexity of getting a non-trivial system to work.
- 4 **Quality of results.** Whether the algorithm is rigorously demonstrated to give good empirical performance on the task considered (here, "real-world" data or "real" experiments may be more effective than "artificial" or "toy" experiments); or whether the theoretical results are strong and interesting; etc.
- 5 **Insight conveyed.** Whether the paper conveys insight into the nature of an algorithm; into the nature of a practical application or problem; into general lessons learned; and/or into theoretical or mathematical tools that might be used by others for future work.

- Not all papers are expected to address all of these criteria, and a paper that is extremely strong on only one of them may well be acceptable for publication.
- For example, a learning theory paper that studies an existing algorithm may be reasonably expected to address only the last of these criteria.
- However, in some cases where the research can be reasonably expected to address more than one of the criteria above, a paper may have a better chance of acceptance if it does indeed address them.
- For example, a paper that gives an elegant mathematical derivation of a new algorithm (Criterion #1) may fare better if it is also demonstrated through rigorous empirical evaluation to do well (Criterion #4), or demonstrated on a real/non-trivial application (Criterion #3). This is because such experiments can help build a significantly stronger case for the algorithm's actual utility.
- Similarly, a paper describing an impressive application of machine learning (Criterion #2 or #3) may fare better if beyond reporting success, it further elucidates the structure of the problem or algorithm that made the application work, and thereby conveys insight (Criterion #5).
- For empirical studies, a good result can lie along many different axes, all of which compare to the best state-of-the-art algorithm. These axes may include: better accuracy, better ROC performance, faster, less memory, more generally applicable, easier out-of-the-box usage, much simpler to code. If an algorithm does not excel along any of these axes, a reviewer may wonder why it is worth publishing at NIPS.
- Although NIPS strongly encourages interdisciplinary work that spans multiple topics, we now also describe some evaluation criteria that are more specialized and may apply only to individual topics.

- e.g., clustering, dimensionality reduction, feature selection, nonparametric Bayesian models, graphical models, kernels, boosting, Monte Carlo methods, neural networks, semi-supervised learning, deep learning.
- Authors of papers that propose new algorithms for well-established, existing problems are encouraged to provide evidence for the practical applicability of their methods, such as through rigorous empirical evaluation of their methods on real data or on real problems.
- For example, a paper about a new mathematical trick (or about a beautiful new mathematical derivation) would be stronger if it is supported by empirical evidence that the resulting algorithm really helps on a practical real-world problem.
- NIPS also encourages submission of papers that describe algorithmic or implementation principles that may have a large impact on applications or on practitioners of machine learning.

- Authors of papers that propose new algorithms for existing problems (such as solving MDPs) are encouraged to provide rigorous empirical evaluation of their methods on real-world problems, and show its relevance to **real/difficult decision making** or control tasks.
- For example, rather than demonstrating your idea only on a grid-world or on mountain-car, also show if it works on a more challenging task. The other comments for algorithmic papers also apply here.
- Learning Theory
- Any Learning Theory paper should have a theorem about learning and a proof. Leaving out the proof is not an option in a double-blind setting! Several styles of papers exist:
 - 1 Propose a new natural model of learning and algorithm for this model (examples: Bayes learning, statistical learning, PAC learning, Online learning, MDP learning, Boosting).
 - 2 Propose an algorithm with an improved analysis in some standard setting.
 - 3 Prove that some learning task people have been attempting is hard or impossible.
 - 4 "Other". Meta-theorems about learning theorems, etc. Technically difficulty or novelty is not the goal. Impact on the process and practice of learning is the goal. Experimental results are nice but not necessary in general.

- Application papers should describe your work on a "real-world" as opposed to "hypothetical" application;
- specifically, it should describe work that has direct relevance to, and addresses the full complexity of, solving a non-trivial problem.
- Authors are also encouraged to convey insight about the problem, algorithms, and/or application. For example, one might describe the more general lessons learned, or elucidate (through an ablative analysis/lesion analysis, which removes one component of an algorithm at a time) which were the key components of the system needed to get the application to work.
- A NIPS application paper should be comparable in quality to paper in the corresponding application domain conference: for example, a text paper should be acceptable to SIGIR, EMNLP, or other appropriate conference. Application papers should not only present concrete application results, but also contain at least one of the below elements:
 - Applications that couldn't previously be done, at all, or on this scale
 - Techniques shown to be uniquely fitted to specific popular applications, leading to improved performance or more accurate solutions
 - Insights that, from the perspective of machine learning, distinct applications X and Y, whose respective users have never talked to each other, are the same.
 - Careful analytic studies that may not demonstrate improved performance but that compare different approaches on large representative corpora.
 - Such studies should provide insights e.g. regarding performance gains achievable by using more complex learning machines, and the relative importance of preprocessing and feature selection.

- Authors of vision papers are encouraged to provide rigorous empirical evaluation of their methods to demonstrate value added not just for a few selected images, but more broadly.
- Ideally, a NIPS paper proposes a machine learning algorithm or system that can be used by a computer vision researcher to help solve a difficult computer vision problem.
- NIPS papers in this area should be comparable in quality to those accepted in the major computer vision conferences, such as ICCV or CVPR.
- Speech and Signal Processing
- Similar to computer vision, a NIPS paper should solve a difficult audio, speech, or other signal processing problem via machine learning; and be useful for a signal processing practitioner. The quality bar for NIPS is higher than those of a typical signal processing conference (such as ICASSP or ICIP): the NIPS papers are 30% longer, the reviews are more detailed, and the acceptance rate is about half. Therefore, a NIPS signal processing paper should be more significant than the average ICASSP paper.
- Hardware Technology
- In addition to describing a successful implementation, a NIPS hardware paper should also convey insight into the underlying principles behind your implementation that serve as useful lessons learned to non-hardware researchers, such as computer scientists or neurobiologists.

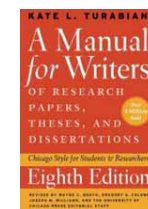
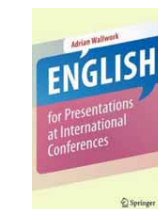
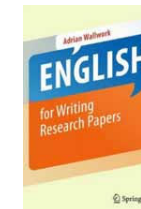
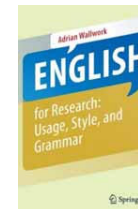
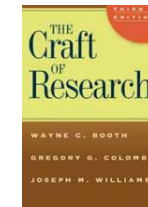
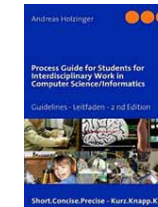
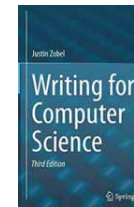
- A significant fraction of NIPS papers, comprising mainly ones from the neuroscience, biological vision, or cognitive science, either describe or study natural systems. Examples include a paper proposing a new model of human decision making, a paper describing evidence for a neural code, and so on. Papers submitted in this category should make significant contributions to the computational, psychological and/or neural understanding of an important biological and/or behavioral system or function. Such papers will be evaluated on the basis of some or all of the following seven criteria:
- 1 Novelty of model. For example, a new account of a popular issue such as the representation of uncertainty in neural population codes.
- 2 Novelty of method. For example, a new analytical analysis of a phenomenon (say phase locking in oscillatory networks) that had previously only been studied using simulations.
- 3 Novelty of results. For example, a re-analysis of data on input-output functions of auditory cortical neurons, showing a new facet of their tuning to spectral contrast.
- 4 Novelty of system or function. For example, a model of a neural region (say a hypothalamic nucleus) that has not hitherto been analyzed.
- 5 Fit to data. For example, whether the suggestion evidently accounts for a wide range of data that have resisted previous approaches.
- 6 Explanatory power. For example, whether the suggestion links different (Marrian) levels of analysis, maybe showing the control-theoretic or Bayesian soundness of a well-known psychological learning rule.
- 7 Appropriateness of model. For example, if a proposed model or mechanism is supported by multiple data points or experiments.
- A good neuroscience model should make testable predictions - and they should be interesting, too. An interesting prediction is something you may not have thought about otherwise: a prediction that is non-obvious, or does not derive directly from the limitation assumptions made in the model.
- A neuroscience model should give you a new way of looking at the system, which inspires new experiments. NIPS neuroscience papers should either be neuro-scientifically or computationally well-grounded, ideally both. The paper should make a serious attempt at connecting to state-of-the-art neurobiology, and/or provide a rigorous mathematical treatment or comparison to a state-of-the-art engineering method.

- Papers on this topic tend to fall between the natural and artificial systems categories. A good brain imaging paper may lead to neurobiological insight, or it may propose an experimental method for obtaining new kinds of measurements. A good brain computer interface would either be useful as a computer interface, or also lead to neurobiological insight.
- These criteria were selected with the goals of encouraging good research, and of maximizing NIPS' long-term impact.
- Note that this is not as simple as accepting papers with high-expected impact.
- For example, a paper that makes ambitious but poorly substantiated claims may have high expected impact---|largely on the off-chance that the claims turn out to be correct---|but is still likely to be rejected. Some of these evaluation criteria exactly address this issue of providing evidence for the utility of one's work.



Thank you!

Appendix



web.efzg.hr/dok/MAT/vkojic/Larrys_speakeasy.pdf

Handbook for
Spoken Mathematics
(Larry's Speakeasy)

Lawrence A. Chang, Ph.D.
With assistance from
Carol M. White
Lisa Abrahamson



HELPFUL: https://en.wikipedia.org/wiki/List_of_mathematical_symbols

LaTeX Symbols : <http://www.artofproblemsolving.com/wiki/index.php/LaTeX:Symbols>

Math ML: <http://www.robinlionheart.com/stds/html4/entities-mathml>

The MathML Association promotes & funds MathML implementations









MathML3 is an ISO/IEC International Standard

- Risk #1: Not achieving the minimum expected scientific output per year
- Mitigation of Risk #1: **No** start of PhD without solid PhD proposal including achievable dissemination milestones and alternatives!
- Risk #2: Recognizing that parts of the PhD goals are unattainable
- Mitigation of Risk #2: Having listed alternative research routes in the PhD proposal and being flexible to change research directions

Google Scholar

Profile

	Carlos Castillo Distinguished Research Professor, Universitat Pompeu Fabra Bestätigte E-Mail-Adresse bei upf.edu Social Computing Social Media Crisis Informatics Algorithmic Fairness	Zitiert von: 15282
	Kai-Wei Chang Assistant professor, UCLA Bestätigte E-Mail-Adresse bei kwchang.net Natural Language Processing Machine Learning Structured Prediction Algorithmic Fairness	Zitiert von: 11080
	Suresh Venkatasubramanian School of Computing, University of Utah Bestätigte E-Mail-Adresse bei cs.utah.edu Algorithmic Fairness Computational Geometry Data Mining Clustering	Zitiert von: 8599
	Arun Sundararajan NYU Stern School of Business Bestätigte E-Mail-Adresse bei stern.nyu.edu network science sharing economy platforms future of work algorithmic fairness	Zitiert von: 7034
	Aaron Roth Associate Professor of Computer Science, University of Pennsylvania Bestätigte E-Mail-Adresse bei cs.upenn.edu Differential Privacy Algorithmic Fairness Algorithmic Game Theory Learning Theory	Zitiert von: 6054
	Bert Huang Assistant Professor, Virginia Tech Department of Computer Science Bestätigte E-Mail-Adresse bei vt.edu Machine learning structured prediction graphical models online harassment detection algorithmic fairness	Zitiert von: 2482