Should we stop doing Deep Learning research in times of AI?

ELLIS Delft Machine Learning Symposium

Nov 19, 2026



Jan van Gemert, TU Delft.
https://jvgemert.github.io/

Warm up questions

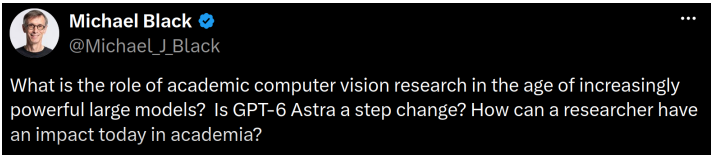
Q: Why did you start a PhD?

Warm up questions

Q: Why did you start a PhD?

Q: Will next month's AI invalidate your research?

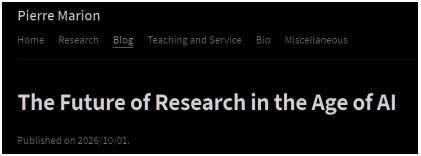
Not alone: standing on the shoulders of giants



https://x.com/Michael_J_Black/status/2097585486466327003 (Sept 9 2026)



https://web.mit.edu/phillipi/www/writing/PhD-in-age-of-AI.html (Sept 28 2026)



https://pierremarion23.github.io/thoughts/2026-10-01.html

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Q: For Deep Learning researchers breaking SOTA for application A, are they themselves often a user of task A? (e.g.: image recognition, world models, image/video generation, ...)

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Q: For Deep Learning researchers breaking SOTA for application A, are they themselves often a user of task A? (e.g.: image recognition, world models, image/video generation, ...)

- No. Breaking SOTA, is solving other people's specific application problems.

(Which is probably why my answer is different from Michael Black's.)

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Q: For Deep Learning researchers breaking SOTA for application A, are they themselves often a user of task A? (e.g.: image recognition, world models, image/video generation, ...)

- No. Breaking SOTA, is solving other people's specific application problems.

(Which is probably why my answer is different from Michael Black's.)

AI competes with application-building for other people's problems.

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Q: For Deep Learning researchers breaking SOTA for application A, are they themselves often a user of task A? (e.g.: image recognition, world models, image/video generation, ...)

- No. Breaking SOTA, is solving other people's specific application problems.

(Which is probably why my answer is different from Michael Black's.)

AI competes with application-building for other people's problems.

- Why are we working on other people's problems SOTA?
- Isn't our own field worthy of scientific investigation?

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Q: For Deep Learning researchers breaking SOTA for application A, are they themselves often a user of task A? (e.g.: image recognition, world models, image/video generation, ...)

- No. Breaking SOTA, is solving other people's specific application problems.

(Which is probably why my answer is different from Michael Black's.)

AI competes with application-building for other people's problems.

- Why are we working on other people's problems SOTA?
- Isn't our own field worthy of scientific investigation?

Q: What are Deep Learning researchers for? (hint: not for building the specific applications)

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Q: For Deep Learning researchers breaking SOTA for application A, are they themselves often a user of task A? (e.g.: image recognition, world models, image/video generation, ...)

- No. Breaking SOTA, is solving other people's specific application problems.

(Which is probably why my answer is different from Michael Black's.)

AI competes with application-building for other people's problems.

- Why are we working on other people's problems SOTA?
- Isn't our own field worthy of scientific investigation?

Q: What are Deep Learning researchers for? (hint: not for building the specific applications)

- Investigate, understand, communicate how to build foundation/frontier AI.

Existential questions

Q: For what applications are Deep Learned foundation/frontier AI models useful?

- AI as a Swiss army knife: Used deeply and broadly in *any* specific application.

Q: For Deep Learning researchers breaking SOTA for application A, are they themselves often a user of task A? (e.g.: image recognition, world models, image/video generation, ...)

- No. Breaking SOTA, is solving other people's specific application problems.

(Which is probably why my answer is different from Michael Black's.)

AI competes with application-building for other people's problems.

- Why are we working on other people's problems SOTA?
- Isn't our own field worthy of scientific investigation?

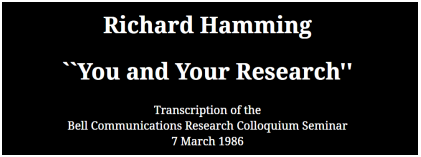
Q: What are Deep Learning researchers for? (hint: not for building the specific applications)

- Investigate, understand, communicate how to build foundation/frontier AI.

Work on things that *you* find meaningful ie: (interesting/important/..)

Shoulders of older giants

Exercise: replace his term "computers" with "AI"; "door open/closed" with "working from home" in his talk.



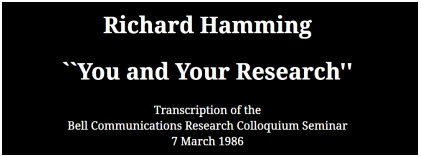


https://www.cs.virginia.edu/~robins/YouAndYourResearch.html

https://www.johndcook.com/blog/2013/02/14/you-and-your-research/

Shoulders of older giants

Exercise: replace his term "computers" with "AI"; "door open/closed" with "working from home" in his talk.





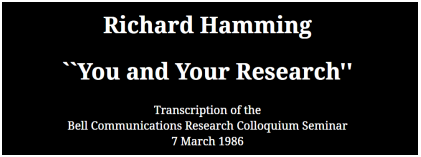
https://www.cs.virginia.edu/~robins/YouAndYourResearch.html

https://www.johndcook.com/blog/2013/02/14/you-and-your-research/

"What are the important problems of your field?"

Shoulders of older giants

Exercise: replace his term "computers" with "AI"; "door open/closed" with "working from home" in his talk.





https://www.cs.virginia.edu/~robins/YouAndYourResearch.html

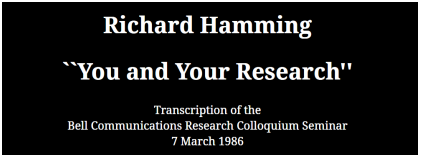
https://www.johndcook.com/blog/2013/02/14/you-and-your-research/

"What are the important problems of your field?"

"If what you are doing is not important, and if you don't think it is going to lead to something important, why are you working on it?"

Shoulders of older giants

Exercise: replace his term "computers" with "AI"; "door open/closed" with "working from home" in his talk.





https://www.cs.virginia.edu/~robins/YouAndYourResearch.html

https://www.johndcook.com/blog/2013/02/14/you-and-your-research/

"What are the important problems of your field?"

"If what you are doing is not important, and if you don't think it is going to lead to something important, why are you working on it?"

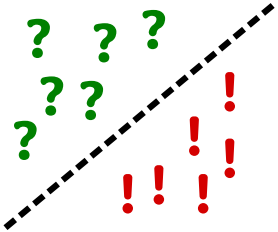
"If you do not work on an important problem, it's unlikely you'll do important work."

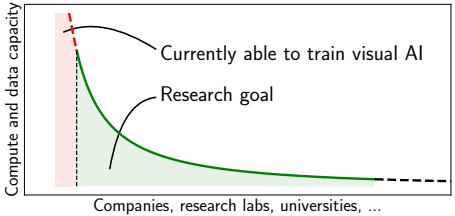
WHOAMI; and my answer.



Jan van Gemert
Computer Vision lab @ TU Delft.

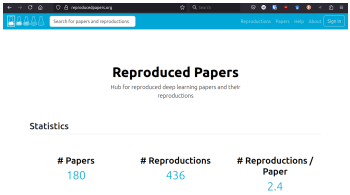
Two main research themes:



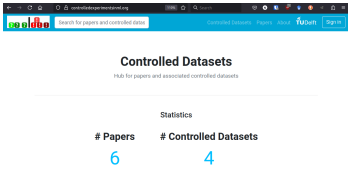


Fundamental empirical understanding-based deep learning research; (to) Design data-efficiency visual AI models; allowing for training outside big tech.

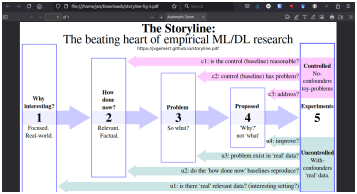
Facilitating fundamental empirical research in ML/DL



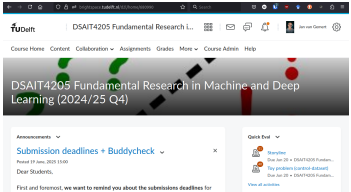
ReproducedPapers.org



ControlledExperimentsInML.org

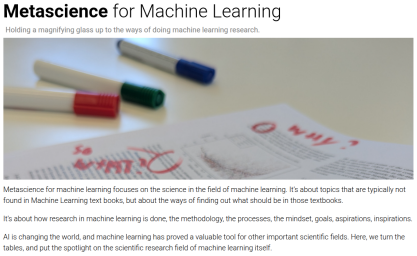


Online research guidelines

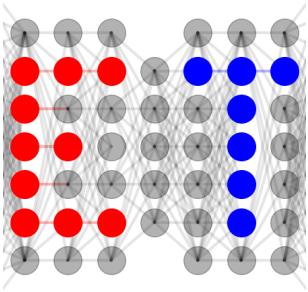


MSc course

Community-building efforts for making a change:



`https://metascienceforml.github.io/`

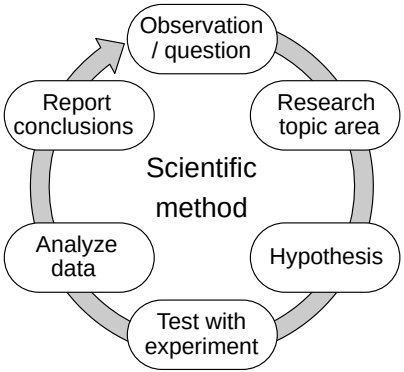


ECCV'26 workshop: E.T.

`https://empiricaltheory.github.io/`

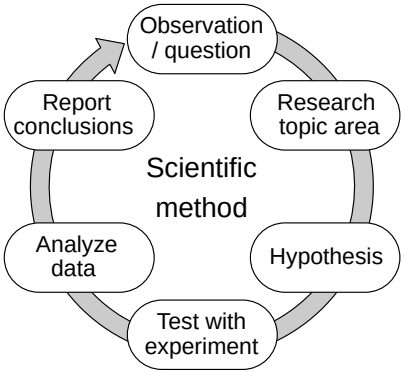
My opinion: Making a change by changing existing structures is more difficult than making a change by giving alternatives. Be the change you want it to be.

Often, deep learning is scientifically[1] shallow [2,3,4]



:

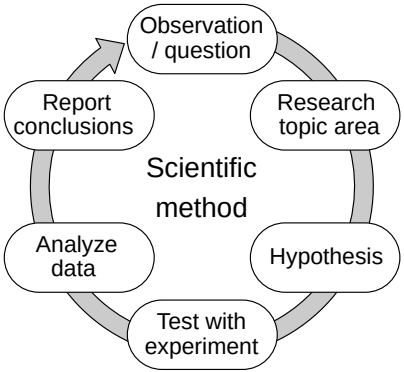
Often, deep learning is scientifically[1] shallow [2,3,4]



:

- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

Often, deep learning is scientifically[1] shallow [2,3,4]

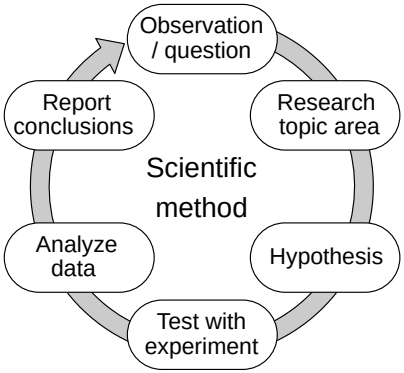


:

- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

Often, deep learning is scientifically[1] shallow [2,3,4]



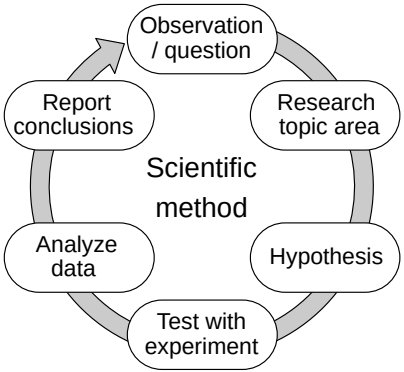
- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

:

"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

- ML/DL has not many empirical theories.

Often, deep learning is scientifically[1] shallow [2,3,4]



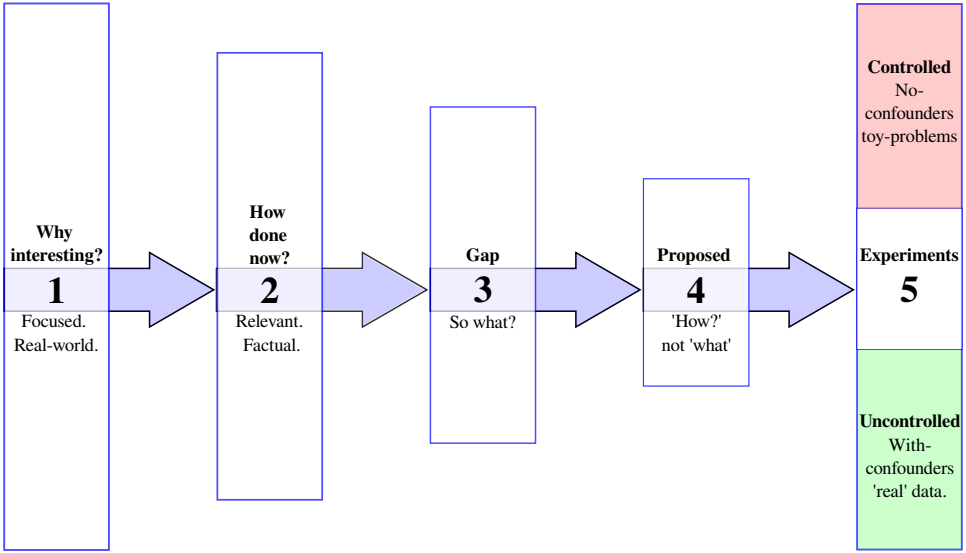
- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

:

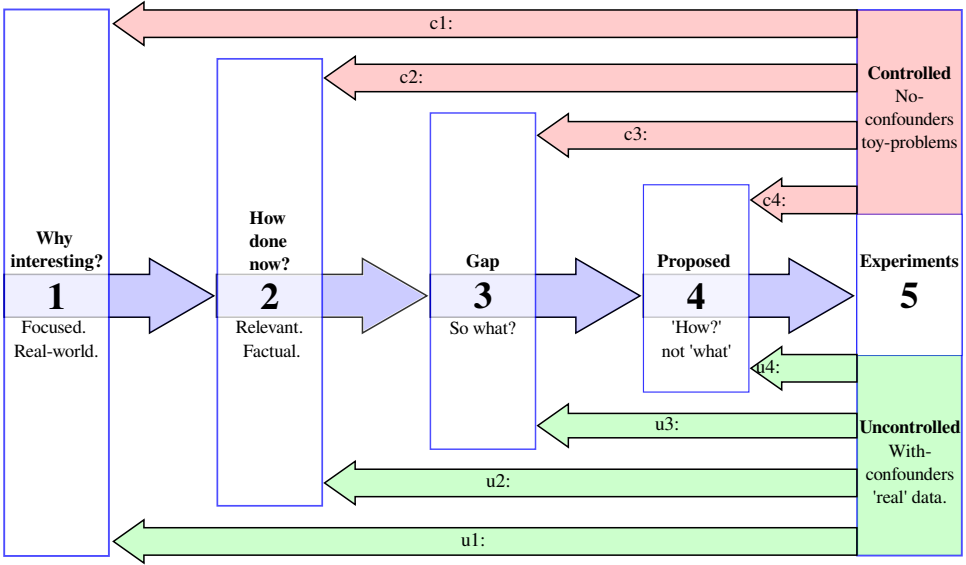
"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

- ML/DL has not many empirical theories. Some that I am aware of:
  - Neural Scaling Laws; ○ Bias/variance ○ ML is like physics/neuroscience;
  - Simple axioms explain intelligence ○ Platonic representation hypothesis ○ ...

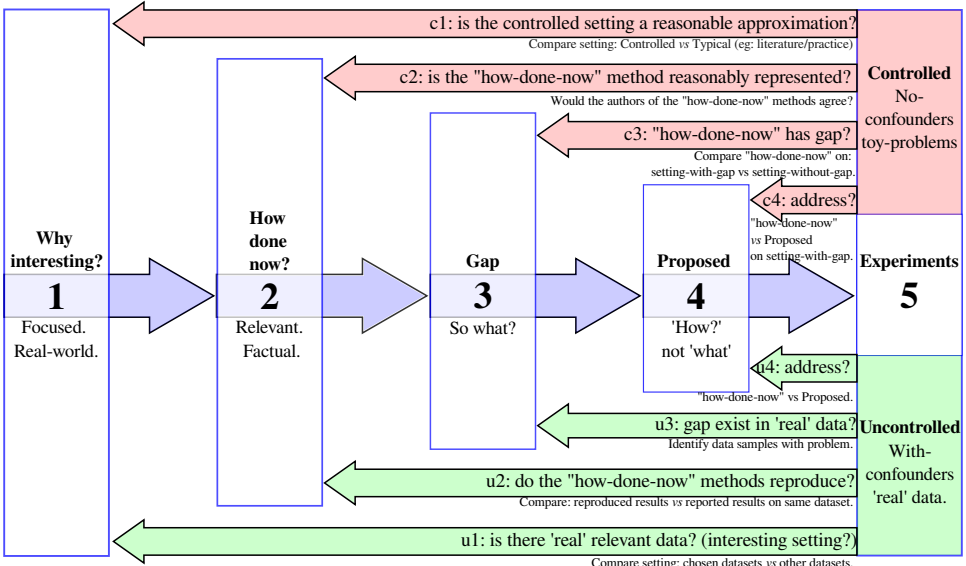
Storyline: Sound argumentation with empirical rigor



Storyline: Force empirical evidence for each step

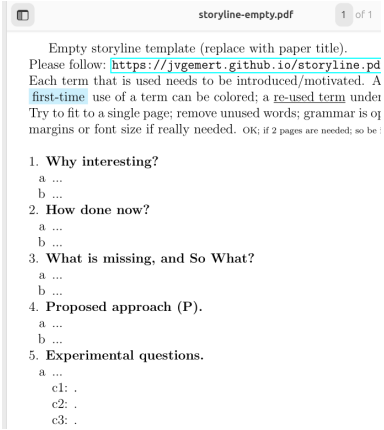


Storyline: Actionable for both author and reviewer



Storyline encourages empirical rigor

- Storyline required to fit on 1 page to have all points available at the same time.
- ~~Forces~~ Helps the authors to provide empirical evidence
- Reviewers can be asked to review the storyline for *argumentation* (point 1-4 in the storyline) and for *Empirical rigor* (the arrows)


storyline-empty.pdf 1 of 1

Empty storyline template (replace with paper title).

Please follow: https://jvgemert.github.io/storyline.pd

Each term that is used needs to be introduced/motivated. A first-time use of a term can be colored; a re-used term under

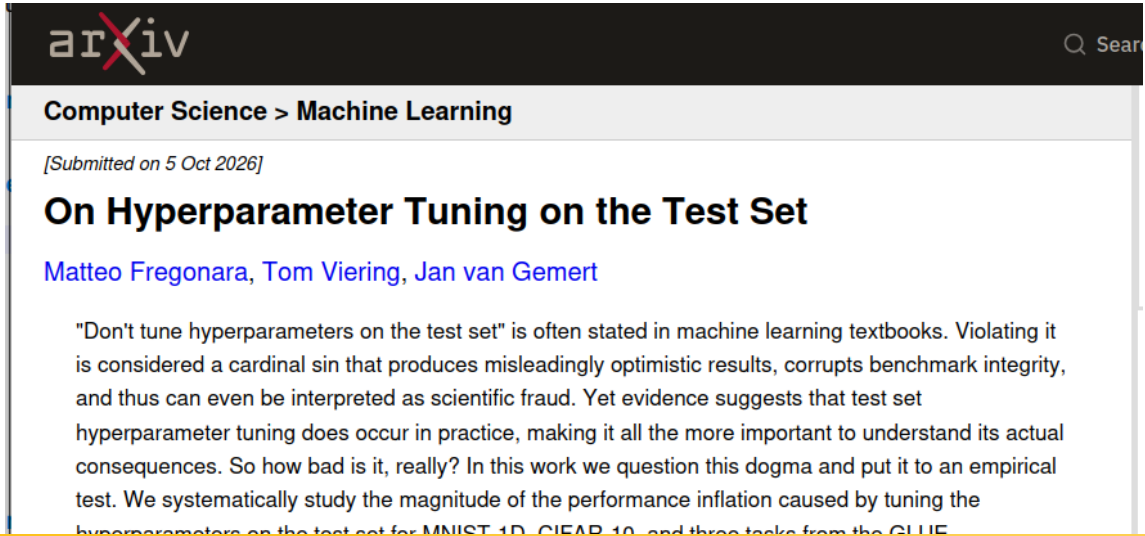
Try to fit to a single page; remove unused words; grammar is op margins or font size if really needed. OK; if 2 pages are needed; so be

1. **Why interesting?**
   a ...
   b ...
2. **How done now?**
   a ...
   b ...
3. **What is missing, and So What?**
   a ...
   b ...
4. **Proposed approach (P).**
   a ...
   b ...
5. **Experimental questions.**
   a ...
   c1: .
   c2: .
   c3: .

https://jvgemert.github.io/storyline.pdf

Evidence 1: My own (perceived) meaningful research

https://arxiv.org/abs/2610.05902



arXiv

Search

Computer Science > Machine Learning

[Submitted on 5 Oct 2026]

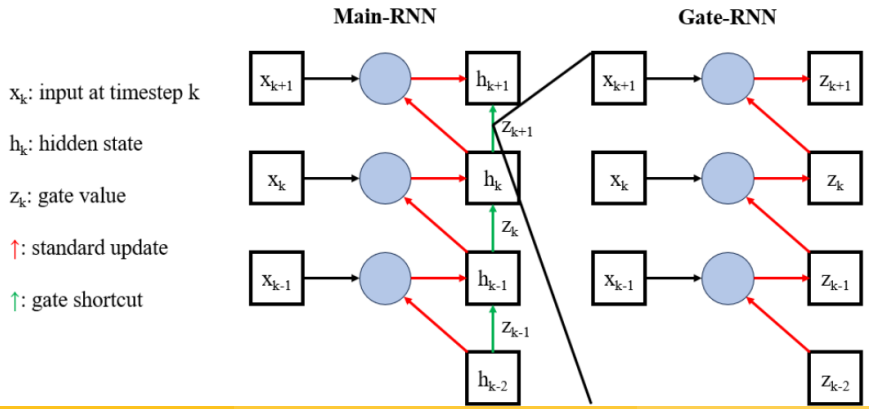
On Hyperparameter Tuning on the Test Set

Matteo Fregonara, Tom Viering, Jan van Gemert

"Don't tune hyperparameters on the test set" is often stated in machine learning textbooks. Violating it is considered a cardinal sin that produces misleadingly optimistic results, corrupts benchmark integrity, and thus can even be interpreted as scientific fraud. Yet evidence suggests that test set hyperparameter tuning does occur in practice, making it all the more important to understand its actual consequences. So how bad is it, really? In this work we question this dogma and put it to an empirical test. We systematically study the magnitude of the performance inflation caused by tuning the

Evidence 2: My own (perceived) meaningful research

https://arxiv.org/abs/2610.05902



Evidence 3: My own (perceived) meaningful research

https://arxiv.org/abs/2610.05902

MODELING SPATIAL TRANSFORMATIONS INSTEAD OF MEMORIZING TRANSFORMED APPEARANCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning spatial transformations from transformed images requires surprisingly large amounts of data and model capacity. We study this effect using controlled affine transformations of MNIST and examine how it changes with dataset size, model capacity, and generalization to unseen appearance and transformation combinations. We evaluate several existing classifier families, including Vision Transformers, ConvNeXts, Spatial Transformer Networks, and group-equivariant CNNs, spanning substantially different spatial inductive biases. We propose a

Conclusion

- Aren't the research questions in our own field meaningful?
- What are the most meaningful questions in our field (and are we working on it?)

Conclusion

- Aren't the research questions in our own field meaningful?
- What are the most meaningful questions in our field (and are we working on it?)

Examples of what I find meaningful:

- Taking action to make our field do rigorous empirical research.
- Methods that generalize beyond a single system. (fundamental research)
- Working on methods that make AI training less dependent on big-tech.

Conclusion

- Aren't the research questions in our own field meaningful?
- What are the most meaningful questions in our field (and are we working on it?)

Examples of what I find meaningful:

- Taking action to make our field do rigorous empirical research.
- Methods that generalize beyond a single system. (fundamental research)
- Working on methods that make AI training less dependent on big-tech.

What do you find meaningful?

Conclusion

- Aren't the research questions in our own field meaningful?
- What are the most meaningful questions in our field (and are we working on it?)

Examples of what I find meaningful:

- Taking action to make our field do rigorous empirical research.
- Methods that generalize beyond a single system. (fundamental research)
- Working on methods that make AI training less dependent on big-tech.

What do you find meaningful?

References:
[1]: https://en.wikipedia.org/wiki/Scientific_method
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.
[3]: Sculley, David, et al. "Winner's curse? On pace, progress, and empirical rigor." 2018.
[4]: Togelius, Julian, et al. "Choose your weapon: Survival strategies for depressed AI academics" Proc. of the IEEE (2024).