An Opinionated Reference on Doing Research

Tossaporn Saengja

Contents

1. Personal Experience

- I'm writing this in a commanding tone to my future self to avoid repeating mistakes.
- Learn to use all leverages:
  - LLM is very useful:
    - Kickstart labor tasks: "Can you write a Python script to find the most similar images `dir1` for each image in `dir2` using L2?"
    - Brainstorm: "What are the best metrics for crafting a good food nutrition dataset? Can you give pseudocode for each?"
- Direction is more important than speed, but both are needed.

1.1. Experience at VLL

- Get more ideas from reading or walking.
- Capture ideas to any form of writing.
- Design experiments with clear purpose and **hypothesis**.
- Make good visualizations.
  - Bad visualizations hide findings.
  - Rearrange visualizations when comparing two things.
  - There are always more ways to visualize data.
- Always do anything that seems worth doing. Don't be lazy.
  - It's okay if it's time-consuming the first time.
  - When it happens repeatedly, automate.
- Always be aware of the process and optimize the bottleneck.
  - Need to rearrange images for research update for the third time? Ask ChatGPT for a Python script to automate it.
  - Difficult to run experiments on different machines? Set up a container.
- Have multi-level perspectives, debug (smallest) to experiment design (highest?).
  - Know when to zoom in and out.
- Fast feedback loop is very important.
  - Set up environments, tools as much as possible to eliminate brain blocks.
    - Ideas slip away while waiting to load a model.
- Introduction is more like an extended abstract.
  - Clear and concise.
  - Easy to understand.
  - Easy to see the contributions.

1.2. Examples

- Visualization
  - spatial
    - bad: two things that are far apart
    - good: reorganize to be close (it's also easier to compare horizontally than vertically)
  - scale
    - bad: two things that are different scales
    - good: normalize/standardize
- Note-taking

- bad: note once and forget
  - you forget about the note because it is not good enough to be a reference for your own
  - personalized it to be YOUR own note.
- good: always refine and update

2. John Schulman

- This[1] is the most solid, high-level advices I have read so far on doing ML research.
- Three rough forms of ultimate research goals
  - groundbreaking result that changed perspective on some problem
  - an algorithmic idea that's reusable
  - a deep insight about some recurring questions
- research taste is important to be developed
- goal-driven to develop unique perspective
  - ask questions that nobody else is asking
- constrain search to solutions that seem general and can be applied to other problems
  - For example, avoid incorporating domain information into the solution–achieve locomotion in simulation, in a general way that could be applied to other problems.
- use notebook
  - review every 1 or 2 weeks
  - read all daily entries
  - condense the information into a summary
  - Usually they contains sections for experimental findings, insights (which might come from yourself, colleagues, or things you read), code progress (what did you implement), and next steps / future work.
- switching problems too frequently (and giving up on promising ideas) is a more common failure mode than not switching enough
- one untested strategy is to devote some fixed time budget to trying out new ideas that diverge from main line of work.
  - one day per week on something totally different
  - This would constitute a kind of epsilon-greedy exploration, and it would also help to broaden knowledge.
- The main ways to build ML knowledge are to read textbooks, theses and papers; and to reimplement algorithms from these sources.
- in early career, it is recommended to split time about evenly between textbooks and papers.
  - choose a small set of relevant textbooks and theses to gradually work through
  - and you should reimplement the models and algorithms from your favorite papers.
  - A couple of John's favorites were Numerical Optimization by Nocedal & Wright, and Elements of Information Theory by Cover & Thomas.
  - Recent theses are often the best place to find a literature review of an active field, but older theses also often contain valuable gems of insight.
  - Textbooks and theses are good for building up your foundational knowledge, but you'll also need to read a lot of papers to bring your knowledge up to the frontier.

[1] http://joschu.net/blog/opinionated-guide-ml-research.html

3. Graham Neubig

- This is noted from "CMU Advanced NLP Fall 2024 (9): Experimental Design and Data Annotation"[2][3].
- Research can be roughly categorized as:
  - Applications-driven: make a better, useful system
  - Curiosity: understand something
  - The ratio from ACL is about 95% applications-driven, 5% curiosity.

3.1. Hypothesis

- "Yes-no" questions are often better than "how to."
- A good hypothesis is explicit, precise, and falsifiable.
  - Certain experiment result can validate or disprove the hypothesis.
  - "Does X make Y better?" is not precise.
  - "Do pre-trained embeddings help more when the size of the training data is small?"

3.2. Data annotation

- Statistically significant difference needs a certain amount of data.
- Given **effect size** and significance threshold, "Power analysis" [1] can estimate the amount.
  - For example, effect size is the expected accuracy difference between tested models.

3.3. Workflow

- Directory can be used in experiment steps modularization.
  - `data/`, `model/`, `result/`, `log/`, `script/`, `note/`
- Name directories by parameters.
  - `transformer-layer8-node512-dropout0.5-labelsmooth0.02`
- Plan results section in advance
  - Create main table, use `TBD` placeholder
  - This helps identify unjustified experimental claims
- Result reporting
  - Generate paper latex directly from log files

4. Simon Peyton Jones

- This is a classic advice on how to write a paper.[4]
- Graham said this is a timeless piece that was useful for him.
- Seven simple, actionable suggestions

4.1. Don't wait: write

- This is like a lazy evaluation.
- Forces us to be clear and focused and crystallises what we don't understand.
- It also makes it easy to share.
- This makes writing part of the research process.

---

[2] https://www.youtube.com/watch?v=hs37ze1j41A

[3] https://phontron.com/class/anlp-fall2024/assets/slides/anlp-09-experimentation.pdf

[4] https://www.microsoft.com/en-us/research/academic-program/write-great-research-paper/

4.2. Identify key idea

- don't need to be the best idea, refine it.
- paper should have **one clear, sharp idea** and be **explicit**
  - "The main idea of this paper is ..."
  - "In this section we present the main contributions of the paper."
- review
  - "I believe the main idea is ..."
- fool-proof the writing

4.3. Tell a story

- imagine explaining at a whiteboard to a friend
  - here's a problem. it's interesting and unsolved.
  - here's my idea
  - it works. compare with others.
- out of 1000 readers on the title, only 3 readers will read the details.

4.4. Nail your contributions to the mast

- don't waste time. state what's interesting. don't state the obvious.
  - bad: "computer programs often have bugs. it is very important to eliminate these bugs ..."
  - good: "consider this program, which has an interesting bug ... We will debug this."
- should be refutable.
- bulleted list
- no "rest of this paper is ... Section 3 provides ..."
  - "We give the syntax and semantics ... (Section 3)"
- page one is very important.
  - is every section referred to from the first page?
  - evidence supports claims from page one

4.5. Related work: later

- it's tiring since it's very compressed.
- be generous to the competition. "in his inspiring paper [...]... We develop his foundation ..."
- acknowledge help from people, and also acknowledge weaknesses in our approach.
- provide value judgement, not just a list of references.

4.6. Put your readers first

- don't send them to sleep or stupidity.
  - if the idea is clever, readers will find it.
- explain with examples, and generalize
- get help, each reader can read for the first time once.

4.7. Listen to your readers

- explain what kind of feedback we want
  - ex. "i got lost here" is much more important than "jarva is mis-spelt"
- "could you help me ensure that I describe your work fairly?"

- treat every review like gold dust.
- read every criticism as a positive suggestion for something we could explain more clearly.
- bad: "you stupid person, I meant X".
- good: fix the paper so that X is apparent even to the stupidest reader.
- thank the reviewers for their time.

5. My Recommended Resources

In the following order:

1. An Opinionated Guide to ML Research[5] (Section 2)
2. How to Read a Paper[6]
3. How to write a great research paper[7] (Section 4)

5.1. Short Opinions

1. How to publish a paper at CVPR[8] gives a good reviewer perspective.
   - Given 70% rejection rate, reviewer's job is to find **ANY** reason to reject the paper
     - "The Cockroach" is a bland paper that is hard to kill.
     - "The Puppy with 6 toes" is a delightful paper that is easy to kill.
2. The Craft of Writing Effectively[9]
   - This is a radical, unconventional, but very useful advices in writing.
   - Write to change ideas, not just explain.
   - Write **VALUES**. Useless pieces won't be read.
   - 50% of PhD time is used to know the readers in the field.
   - Problem + solution is better than background + thesis.
   - Learn the language code
     - Notice what construes *values* in the papers.
3. How to do Research At the MIT AI Lab[10] is outdated but offers good, timeless advices.
   - It paints the setting of good interactions within a lab.
   - Connection helps stay informed on the state-of-the-art.
   - Alan Lakien's "How to Get Control of Your Time and Your Life" is a good book.
4. "Do brains backpropagate" - Geoffrey Hinton[11] is a good talk.
   - Understanding brain makes better AI models.

[5] http://joschu.net/blog/opinionated-guide-ml-research.html
[6] https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf
[7] https://www.microsoft.com/en-us/research/academic-program/write-great-research-paper/
[8] https://billf.mit.edu/sites/default/files/documents/cvprPapers.pdf
[9] https://www.youtube.com/watch?v=vtIzMaLkCaM
[10] https://people.cs.umass.edu/~emery/misc/how-to.pdf
[11] https://www.youtube.com/watch?v=VIRCybGgHts

Bibliography

[1] D. Card, P. Henderson, U. Khandelwal, R. Jia, K. Mahowald, and D. Jurafsky, “With Little Power Comes Great Responsibility.” [Online]. Available: https://arxiv.org/abs/2010.06595°