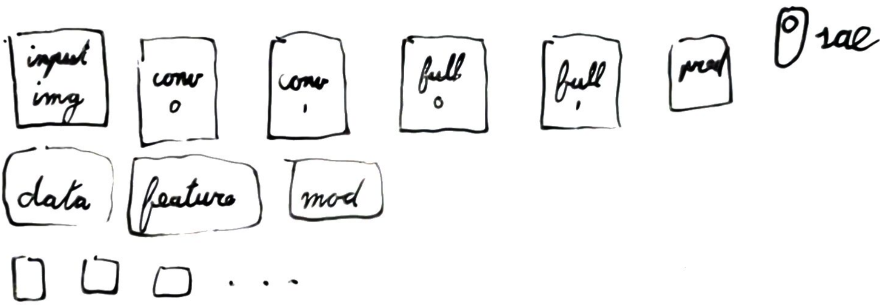
Sparse AutoEncoder



methodology
- new git
- qual. rdk

- quarto bib
- (opt) website

other research
- what problem that they tackle
- (how) are they applied?
- what did they miss?

- nice figures
- maybe empirical comparison

problem
existing solutions
proposed solution
methodology
comparison
implementation
experimental setup
results
discussion, conclusion

problem
solution
methods + implementation
experimental setup
results
comparison
- more results per comparison
conclusion

---

core idea:
interp by id-ness

---

motivation: interp / zoom-in
literature
core idea: interp by id-ness
methodology
experiments
qualitative results + compare lit
quantitative results (ood-ness) + compare benchmarks
conclusion
future work
- mistakes
- what I want to see: more formal id-preserving interp
- mention measuring interp by social science stuff

~~no~~ quantitative results
- reconstruction of the features
- ood analysis of gen imgs
theoretical point of comparison?
- if I fail, jot down possible improvements
after editing diffusion ood benchmarks
how much more ood do I get
compare with reconstructed features