Bugs of the Initial Release of CCS

The Github Repo

https://github.com/collin-burns/discovering_latent_knowledge

- ~~an indention is missing in the training loop~~ (fixed)
- a sigmoid is missing in the linear probe
- an "`.cpu()`" is missing after "`get_first_mask_loc(mask).squeeze()`" (which is by default on the gpu, which is wrong)
- the parser doesn't work (see my implementation)
- `self.probe = self.initialize_probe()` → `self.initialize_probe()`
- No variance normalization by default, though it matters a lot when you have the slightest amount of regularization

The Zip File

See the file like in the Readme of the repo

- For GPT-J, the `appender` should be an empty string rather than a space and an EOS token.
- you have to pass `--datasets all` (the command in the readme doesn't work)
- default is spelt `detaulf` at some place in `extraction_main.py`
- either create some directories or replace the `os.mkdir(directory)` by `os.makedirs(directory, exist_ok=True)`

General Bad Practice

- Normalization is done before the split. This is bad practice and could cause overfitting.
- The scale normalization implemented is not the one described in the paper (it normalize the norm of the vectors, it's not elementwise normalization).
- Figure 1 of the paper is done with the train set.
- The probe initialization is weird: it uses a random spherical initialization, but where the bias is also part of the vector.