Zelda Rose: a tool for hassle-free training of transformer models

Loïc Grobol[1]
(1) Université Paris Nanterre lgrobol@parisnanterre.fr

So you want to train a ✨ transformer ✨ model.

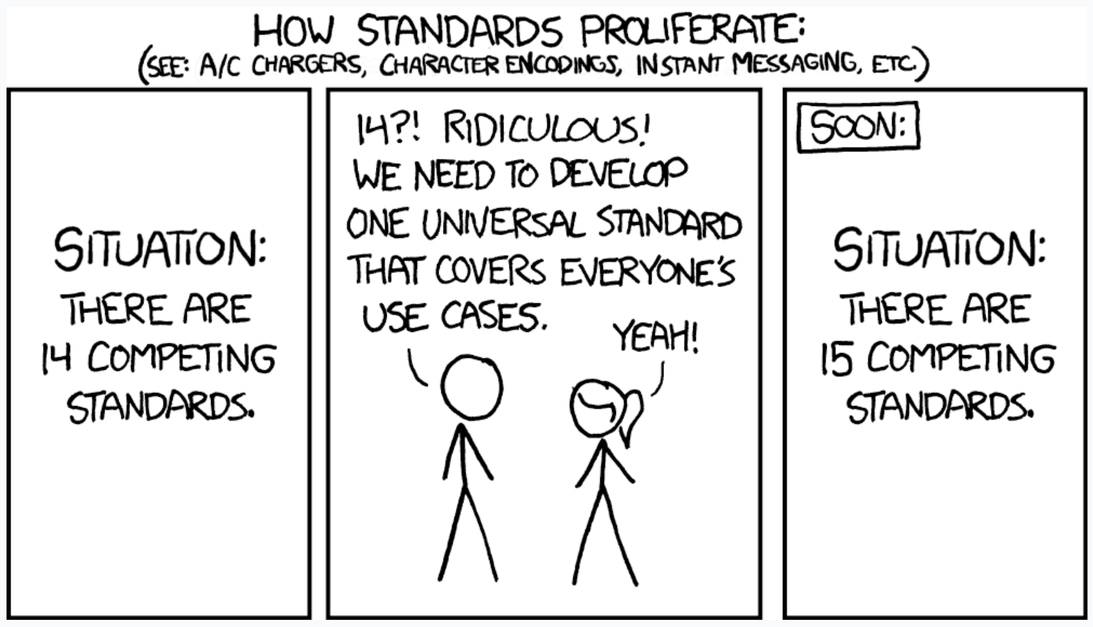
I do!

- You need contextual embeddings for your downstream model, a seq2seq translation… and there are none in your target language/domain/…
- You have painfully gathered a corpus.
- Now what?

Use existing frameworks?

- I really don't want to have to code or to put my data in a weird format.
- There are SO many options.
- Will it run on my cluster?

We need a new ~~framework~~ command-line tool



But I swear this one is good!

So easy to use

Simple CLI invocation:

```
zeldarose transformer \
    --tokenizer roberta_base --model-config roberta_base \
    --val-text dev_corpus.txt \
    train_corpus.txt
```

Simple TOML config:

```toml
type = "mlm"

[task]
change_ratio = 0.15
mask_ratio = 0.8
switch_ratio = 0.1

[tuning]
batch_size = 64
betas = [0.9, 0.98]
epsilon = 1e-8
learning_rate = 1e-4
lr_decay_steps = 1048567
warmup_steps = 1024
weight_decay = 1e-5
```

And that's it!

Just what you need

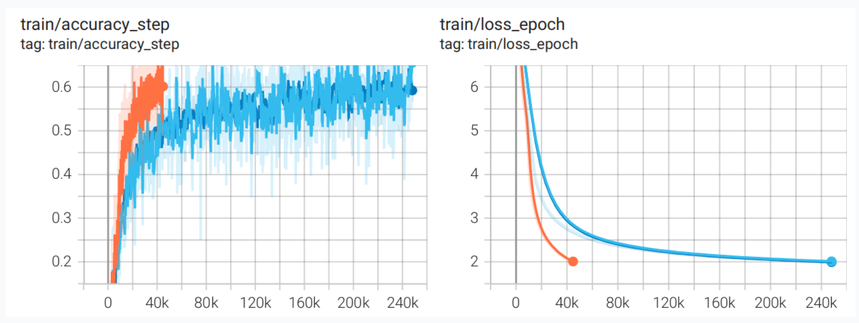
- Train Masked language models and Seq2Seq models.
- New tasks and dataset types are easy to add and maintain.
- Reasonable and battle-tested hyperparameters, configurable as you go.
- We don't do feature creep here.

Library superpowers

We don't reinvent the cheese slicer, we take advantage of cool of-the-shelf laser cutters:

- Load models, configs, tokenizers and datasets from local data or 🤗 hub.
  - Mix and match as you need.
  - Deal with huge datasets using sharding and offloading.
- Run in a plethora of cluster environments and hardware with Lightning ⚡
  - Move to SLURM with just ONE command line flag
  - Run on any number of CPU, GPU, TPU…

No effort feedback



What if the maintainer dies?

- There is a documentation!
- We have (some) tests!
- An automated CI pipeline even!
- The code is REALLY straightforward.
- You can pick it up and make it work by yourself. Promise.

🌹 https://zeldarose.readthedocs.io 🌹