Fast RNN benchmarks

Benchmarks for TorchScript models

For most stable results, do the following:

- Set CPU Governor to performance mode (as opposed to energy save)
- Turn off turbo for all CPUs (assuming Intel CPUs)
- Shield cpus via `cset shield` when running benchmarks.

Some of these scripts accept command line args but most of them do not because I was lazy. They will probably be added sometime in the future, but the default sizes are pretty reasonable.

Test fastrnns (fwd + bwd) correctness

Test the fastrnns benchmarking scripts with the following: `python -m fastrnns.test` or run the test independently: `python -m fastrnns.test --rnns jit`

Run benchmarks

`python -m fastrnns.bench`

should give a good comparison, or you can specify the type of model to run

`python -m fastrnns.bench --rnns cudnn aten jit --group rnns`

Run model profiling, calls nvprof

`python -m fastrnns.profile`

should generate nvprof file for all models somewhere. you can also specify the models to generate nvprof files separately:

`python -m fastrnns.profile --rnns aten jit`

Caveats

Use Linux for the most accurate timing. A lot of these tests only run on CUDA.