On the State of the Art of Evaluation in Neural Language Models

(Melis, Dyer, and Blunsom 2017)

Task: Word Level Language Modelling Datasets: PTB, WikiText-2

Optimizer: [Adam] $\beta_1 = 0$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$ **Batch size**: 64

- Learning rate multiplied by 0.1 whenever validation performance does not improve during 30 consecutive checkpoints (which are performed every 100 and 200 steps for PTB and WT2).

Task: Character Level Language Modelling

- Truncated backprop after 50 timesteps
- Checkpoints every 400 timesteps

Optimizer: [Adam] $\beta_2 = 0.99$, $\epsilon = 10^{-5}$ **Batch size**: 128

Learning to generate reviews and discovering sentiment

(Radford, Jozefowicz, and Sutskever 2017)

Params taken from NVIDIA repro: link

Task: sentiment analysis, transfer learning via a pretrained language model. Datasets: SST, IMDB

- **Model**: 4096-d mLSTM, 64-d embedding, 256-d output. (we also trained a similarly parameterized lstm)
- **Weight Norm**: applied only to lstm parameters (hidden->hidden/gate weights), not embedding or output.
- **Optimizer**: Adam
- **Learning Rate**: 5e-4 per batch of 128. Linear Learning rate decay to 0 over course of epoch.
- **Gradient Clipping**: We occassionally ran into problems with destabilizing gradient explosions. Therfore, we clipped our gradients to a maximum of `1.`.
- **Data set**: Aggressively Deduplicated Amazon Review dataset with 1000/1/1 train/test/validation shards. Each of the three sets are internally shuffled. Samples in a shard are concatenated together so as to persist state across gradient updates.
- **State Persistence**: The hidden state is persisted across all samples and reset at the start of a shard.
- **Batch per gpu**: 128 (instead of OpenAI's 32).
- **Hardware**: 8 volta-class gpus (instead of OpenAI's 4 pascal)
- **Learning Rate Scaling**: We take queues from recent work in training imagenet at scale and leverage FAIR's (Goyal et. al 2017) linear scaling

rule. To account for our 4x batch size increase and 2x gpu increase we used a learning rate of `5e-4 * 8 -> 4e-3`.

Melis, Gábor, Chris Dyer, and Phil Blunsom. 2017. “On the State of the Art of Evaluation in Neural Language Models.” *arXiv Preprint arXiv:1707.05589*.

Radford, Alec, Rafal Jozefowicz, and Ilya Sutskever. 2017. “Learning to Generate Reviews and Discovering Sentiment.” *arXiv Preprint arXiv:1704.01444*.