Meeting — Summary LLM

Bas Donkers, Dennis Fok, Finn-Ole Höner

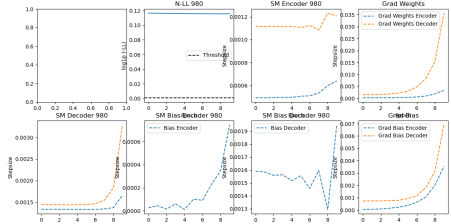
May 13, 2024

Updates

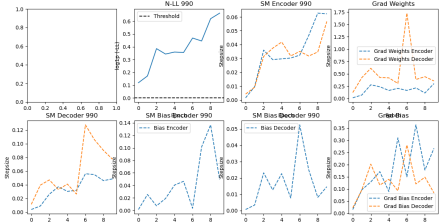
- Gradient spikes first, then large step in decoder / encoder
- These steps seem to affect one hidden unit more strongly than the other
- Increase of Decoder Weight gradients seems to precede the increase of the -LL
- The other gradients seem to react to this jump
- Gradient-Decoder $\rightarrow$ Step $\rightarrow$ Other gradients $\rightarrow$ Step ...
- `torch.nn.utils.clip_grad_norm_(ae.parameters(), max_norm=dMaxGrad)`[1] seems to help. This clips the gradient if the norm exceeds `dMaxGrad`. See here[2].

[1] https://pytorch.org/docs/stable/generated/torch.nn.utils.clip_grad_norm_.html#torch.nn.utils.clip_grad_norm_

[2] https://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/readings/L15%20Exploding%20and%20Vanishing%20Gradients.pdf#page=6.78

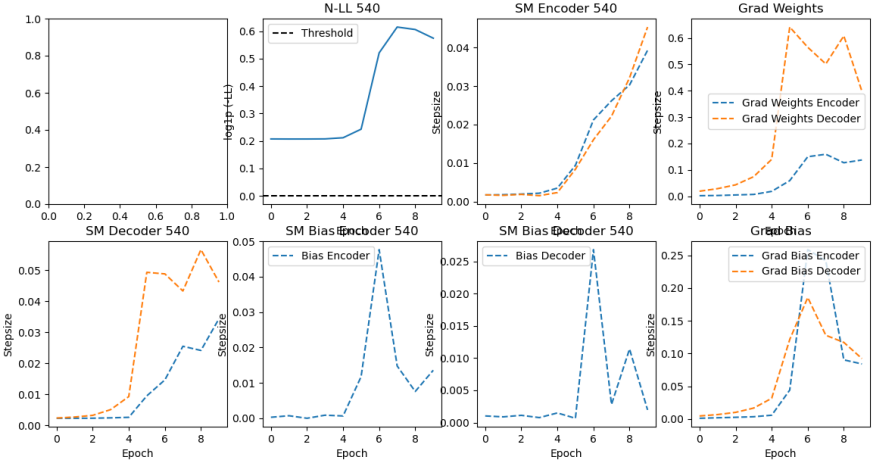


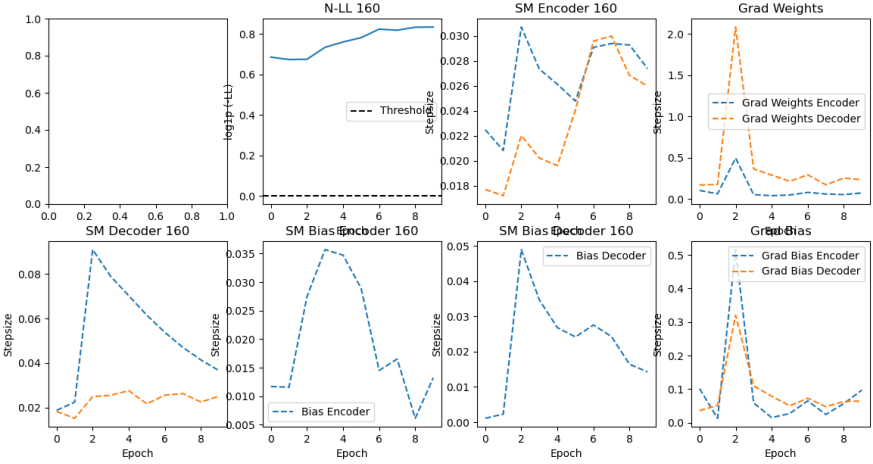
(a) Epoch 970-980

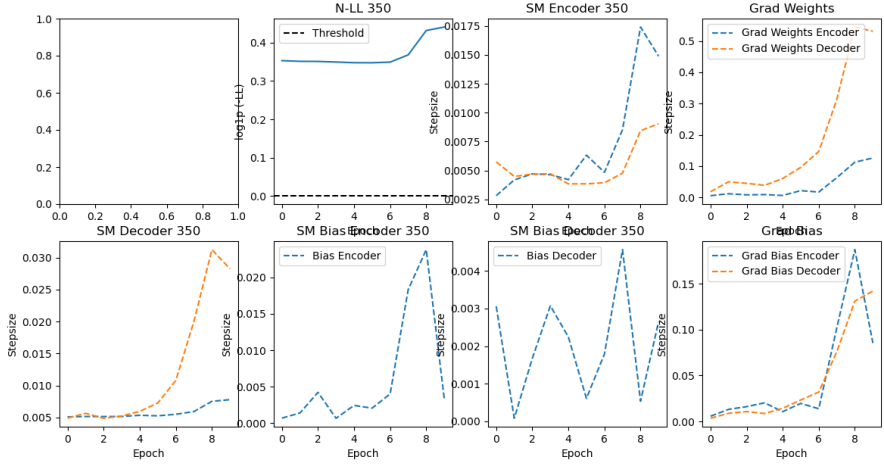


(a) Epoch 980-990

Figure 2: Spikes: Notice how the decoder gradient norm increases first, causing a large step and in-turn affecting the other gradients. The scales of the y-axis differ across all plots.


N-LL 540
Threshold
log1p (-LL)
SM Encoder 540
Stepsize
Grad Weights
Grad Weights Encoder
Grad Weights Decoder
SM Decoder 540
SM Bias Encoder 540
Bias Encoder
SM Bias Decoder 540
Bias Decoder
Grad Bias
Grad Bias Encoder
Grad Bias Decoder
Epoch


N-LL 160
log1p (-LL)
Threshold
SM Encoder 160
Stepsize
Grad Weights
Grad Weights Encoder
Grad Weights Decoder
SM Decoder 160
SM Bias Encoder 160
Bias Encoder
SM Bias Decoder 160
Bias Decoder
Grad Bias
Grad Bias Encoder
Grad Bias Decoder
Epoch


N-LL 350
Threshold
log1p (-LL)
SM Encoder 350
Stepsize
Grad Weights
Grad Weights Encoder
Grad Weights Decoder
SM Decoder 350
Epoch
SM Bias Encoder 350
Bias Encoder
SM Bias Decoder 350
Bias Decoder
Grad Bias
Grad Bias Encoder
Grad Bias Decoder

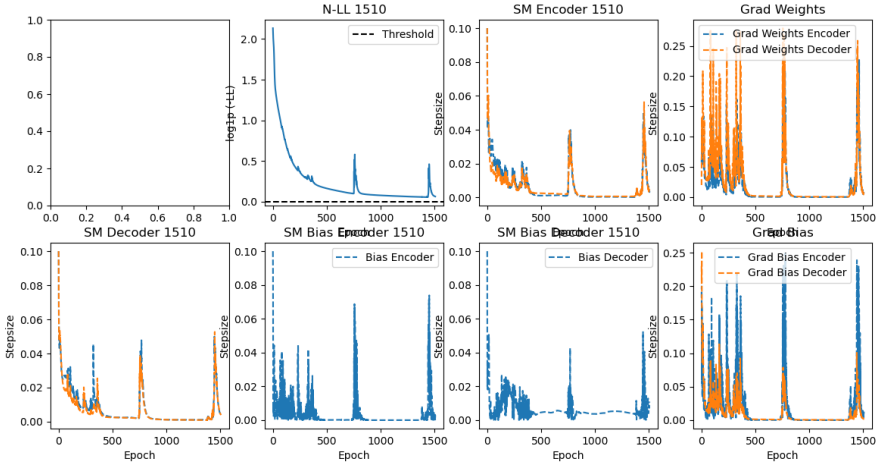


Figure 3: Fixed clipping at 0.3. Later in the optimization this is too large.