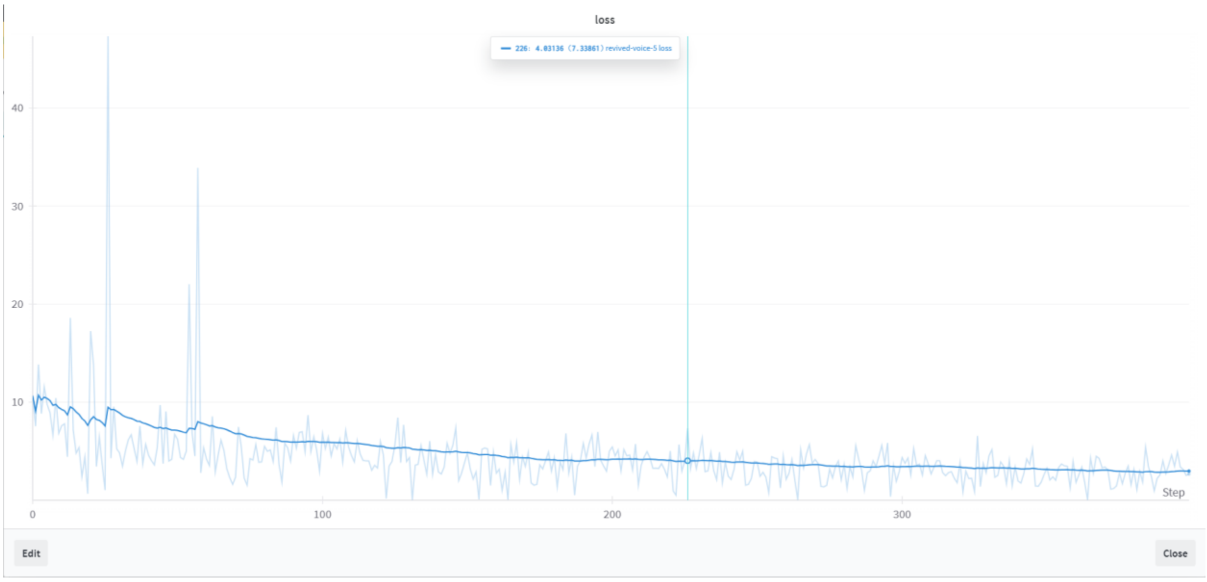
WARP

In my work, I implemented WARP and tested its performance.

I used the lvwerra/gpt2-imdb dataset with movie reviews. First, I trained a reward model by fine-tuning DistilBERT on these reviews. Then, I implemented the WARP method itself and conducted experiments with some hyperparameters.

Overall, it is clear that WARP works. The reward of the baseline model is around 0, but during training, after 200 iterations, the reward increases to 6, and the KL divergence reaches 10. The more steps we take, the more the reward and KL divergence increase, which is logical.

The experiments with the number of steps can be viewed in the exp_results.ipynb file. The training curve of the WARP process looks like this:


loss
226: 4.03136 (7.33861) revived-voice-5 loss
Step
Edit
Close