Summary of Accuracy-Gated vs. Non-Gated Findings

Aaron Pache

ENERGY VS. ACCURACY/ERROR

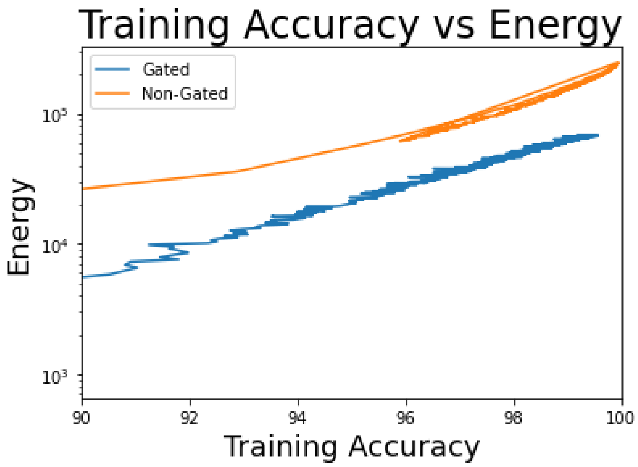

FIG. 1. In training, the gated version provides reasonable energy saving over the non-gated version. The non-gated version, makes a sudden jump in accuracy from 99% to 96% and then climbs back up to 99% accuracy.

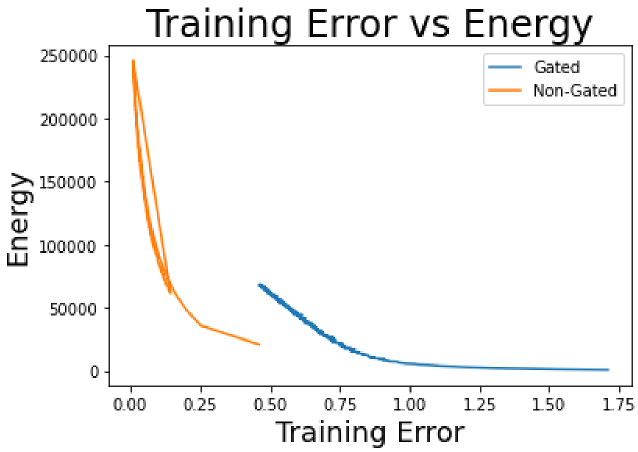

FIG. 2. As the training error gets smaller, energy rises exponentially. I'm not sure if this would be the case for the gated version, since it never reaches a small enough error. Here, the accuracy jump from 99% to 96% in the non-gated version is visible through its training error. Later plots demonstrate this jump more clearly.

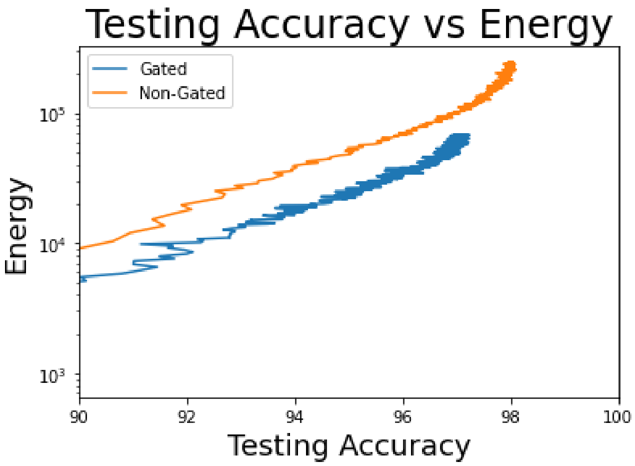

FIG. 3. At high accuracies, energy increases exponentially. The gated version begins to rise in energy at around 97% but the non-gated version rises at around 98%. Even still, the gated version sees energy-saving benefits.

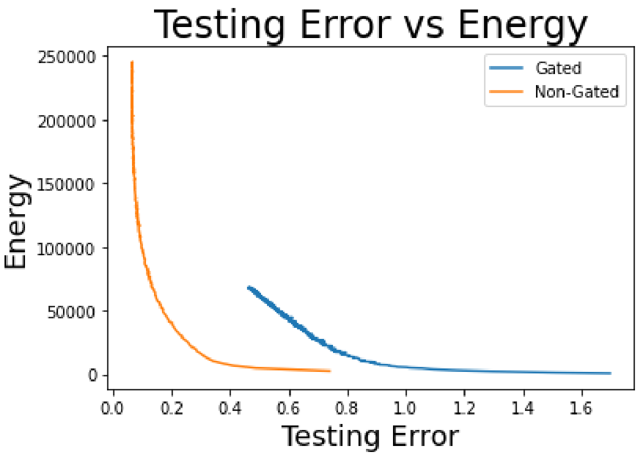

FIG. 4. A similar, if not neater graph to its training counterpart (Figure 2). At small errors the energy increases exponentially and slightly more dramatically compared to the training version. Again, it's difficult to say whether this is applicable to the gated version since it never reaches a low enough accuracy.

MINIMUM ENERGY VS. ACCURACY/ERROR

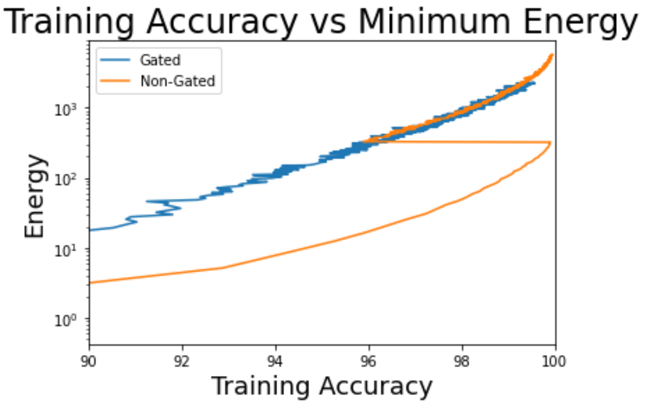

FIG. 5. At minimum energy, the non-gated version consumes less energy. This probably suggests that while it *could* find higher accuracies at lower energies, it doesn't and is therefore more inefficient, atleast in the training case. After this however, it traces the minimum energy of the gated version. I'm not sure if that means that they're taking the same path since the errors are so different...

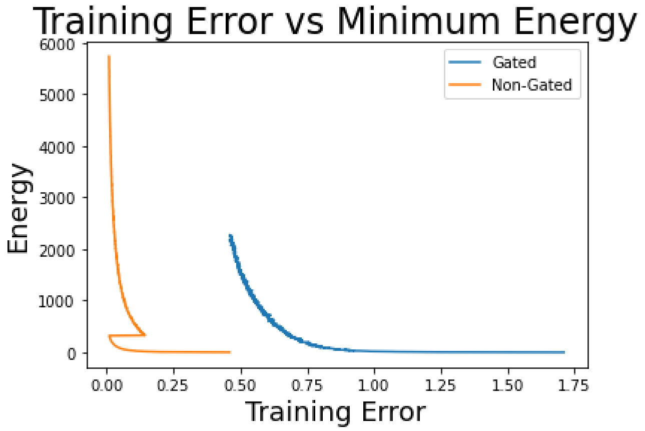

FIG. 6. Unlike the total energy, seen in Figure 2, the minimum energy of the gated version increases exponentially at 0.5 error. While the non-gated version increases at 0.05. My guess is that high accuracies, require high energies but low errors require even higher energies. We infer the accuracy landscape from the error landscape but the error landscape is much more noisy at low errors.

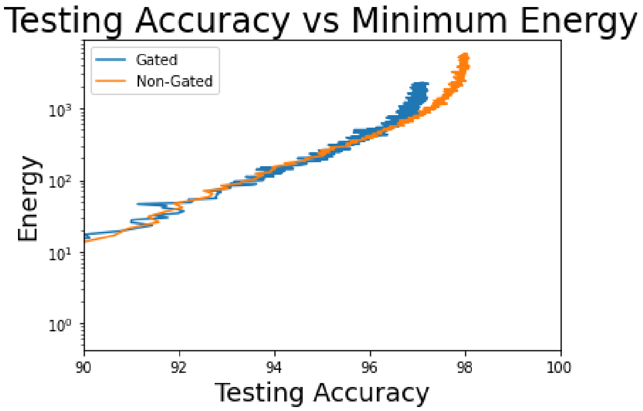

FIG. 7. Under the testing accuracy, both the gated and non-gated appear to walk a similar path until around 96% accuracy. Here, the gated version is able to reach higher accuracies before exponentially increasing in energy. In comparison with the training version (Figure5), it's clear that the non-gated version quickly overfits to the training data before finding a more globally applicable minima.

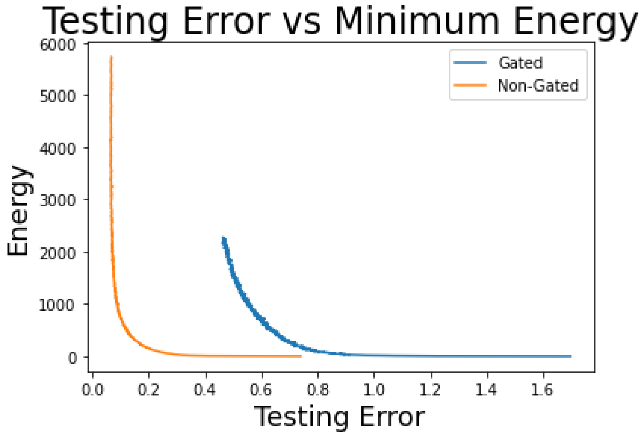

FIG. 8. Again, like training error, the test error energy also increases exponentially for both the gated and non-gated version.

TRAINING ERRORS & ACCURACIES VS. TESTING ERRORS & ACCURACIES

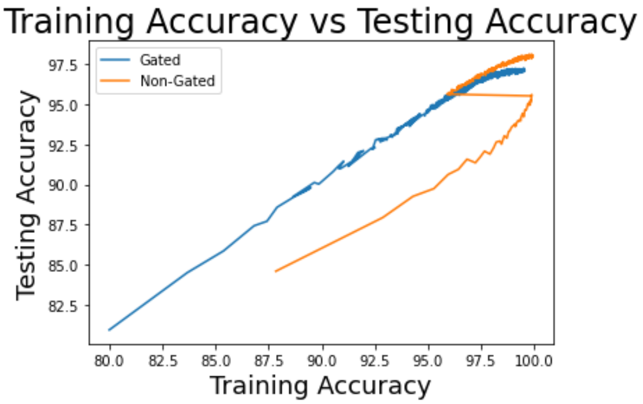

FIG. 9. The gated version follows a more linear path between testing and training accuracies but begins to overfit the training set at 97% accuracy. The non-gated version is overfit from the start before finding a better gradient to traverse where it achieves a better testing accuracy than the gated version.

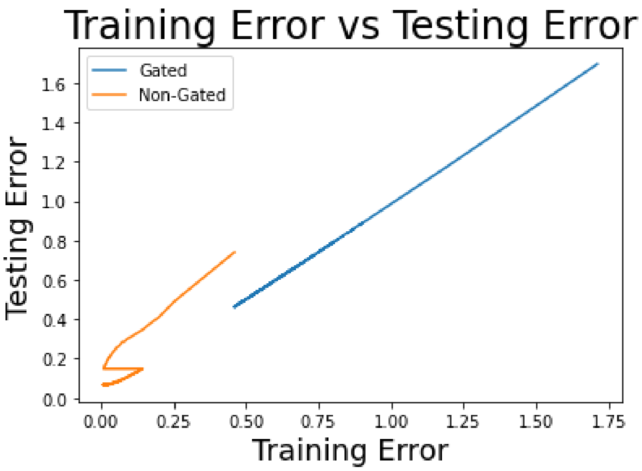

FIG. 10. The gated version does not overfit its training error and is surprisingly linear. The non-gated version overfits the training set error at around 0.1 error.

AN INTERMISSION: MY THOUGHTS

As I allude to in Figure 6, we're interested in the global accuracy landscape and we infer the training accuracy landscape from the training error landscape. I think that at high errors, the training/testing error landscapes are similar and smooth. As the training error decreases, past 0.3, the error landscape is much more noisy and therefore the training and testing errors begin to deviate. The gated version offers a regularisation effect to mitigate falling into this noisy mess by only correcting when necessary (i.e. incorrect samples), this traverses the error landscape much more slowly.

However, the error landscape is used to infer the accuracy landscape. Backpropagation provides a neat method of getting high accuracies following the assumption that 0 error = 100% accuracy. However, it should be clear that the accuracy landscape is not equal to the error landscape. The error landscape contains the noise of every sample which is made explicit at lower errors. The accuracy landscape cuts through the error landscape at an angle and without noise (maybe low-pass filter?). Similar accuracies, but different errors between the gated and non-gated version exist because they lie on the same accuracy surface which cuts through the error surface, Figure 11 attempts to explain this with 1000 words. Unnecessary energy is expended tracking the error landscape of every sample and the least amount of energy likely traces the smoothest error gradient that cuts through each accuracy plane the fastest. Here, I'm imagining that each accuracy plane is like a slice at 90.0%, 90.1%, etc. (i.e. minimise error/maximise accuracy, this way little changes are made to the weights for the most amount of accuracy).

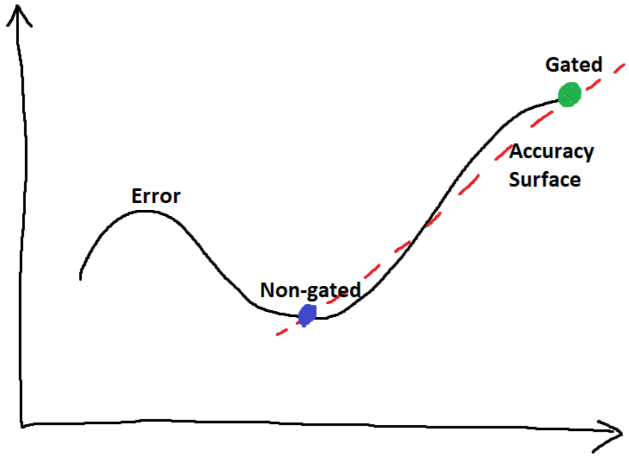

FIG. 11. The error landscape of a given sample. The blue dot represents the non-gated version's position on the error landscape and the green dot represents the gated version's position. The dotted red line represents points where the sample is correctly identified. (I feel like it's at the wrong angle but hopefully the point is conveyed). Energy has been expended bringing the blue ball down the hill, only for this minima to be destroyed on the next sample.

The analogy I internally use is that we each have our own unique handwriting because we lie on the 100% accuracy plane not the 0 error plane. I think a teacher would expend much more energy explaining every detail on how a student should perfectly copy the teacher's 'a'.

SAMPLES VS. ACCURACY/ERROR & TRAINING/TESTING

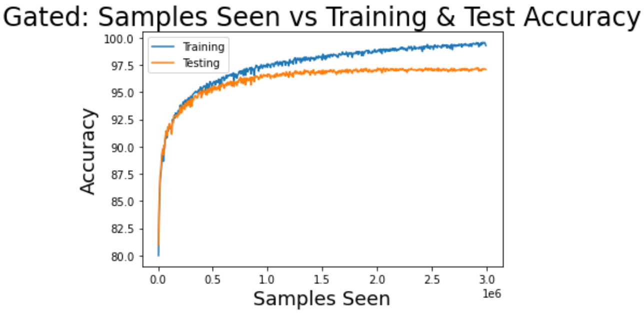

FIG. 12. The gated version begins to overfit the training set accuracy at around 96% accuracy and is able to achieve this accuracy after 750,000 training samples or 12.5 epochs.

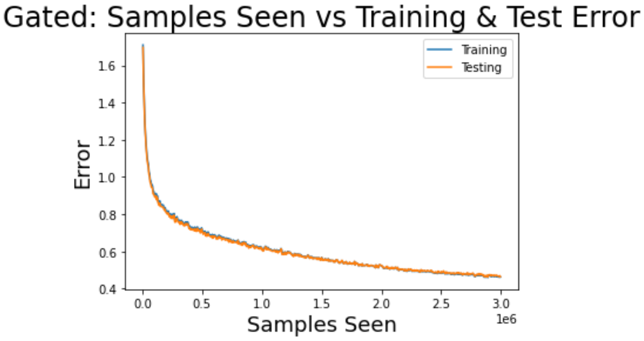

FIG. 13. The gated version does not overfit to the training error landscape, both remain equal for all samples seen and trained against.

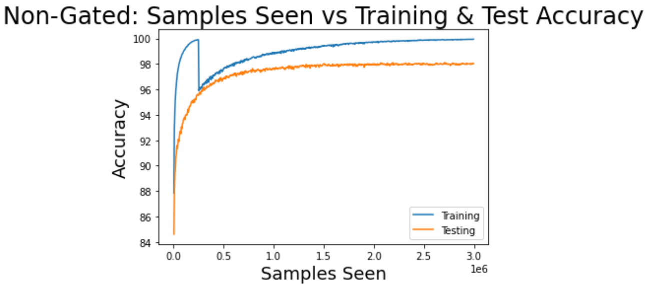

FIG. 14. The non-gated version overfits rapidly, corrects and begins to overfit again. It achieves 96% accuracy at around 250,000 samples or 4.1 epochs.

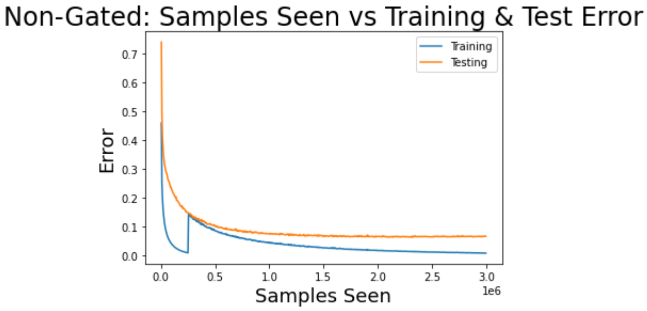

FIG. 15. The non-gated version also overfits the training set's error landscape, unlike the gated version (Figure 13).

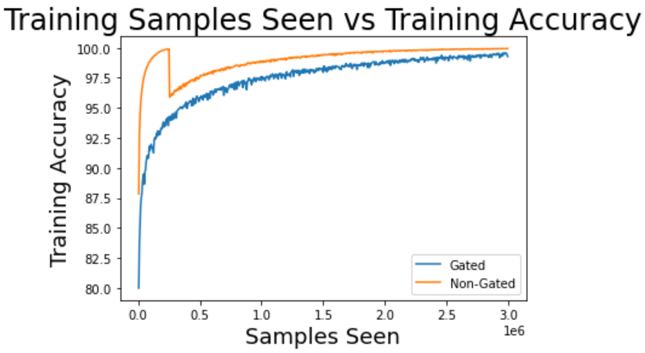

FIG. 16. The non-gated version is able to reach a higher accuracy with less samples seen, 96% accuracy is achieved with 250,000 samples while the gated version takes 750,000 samples.

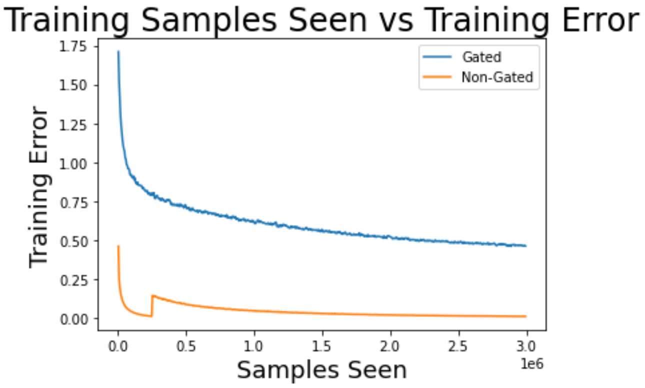

FIG. 17. Again, the non-gated version is capable of reaching a low error with little samples. The gated version prevents the error from getting too low and doesn't saturate as quickly.

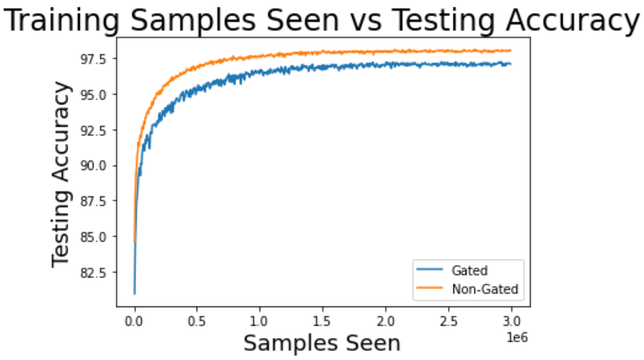

FIG. 18. The non-gated version is able to achieve a higher testing accuracy than the gated version and has a faster rise time. Even though, from Figure 16, the gated version increases in training accuracy, this does not translate to testing accuracy suggesting it has overfit to the training accuracy plane.

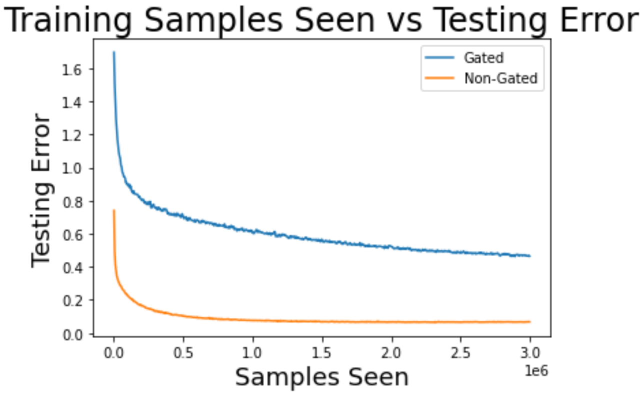

FIG. 19. A similar plot to Figure 17, the training error version and a similar description would be given here too.

I think these plots further corroborate that the training/testing error surfaces are smooth and similar at errors greater than 0.4. From Figures 13, 15, 17 and 19, the errors are similar above 0.4 and overfitting occurs under 0.4. (although, the non-gated version quickly drops below this, so I'm unsure how applicable this extrapolation is).

I think that they also show the global accuracy landscape and global error landscape are similar above 0.75, which is when the gated version begin to deviate in Figure 12. I'm not sure how similar they are beyond these points, but I think the accuracy landscape is smoother than the error landscape beyond these points.

I think the challenge is now: how do we navigate the accuracy surface?