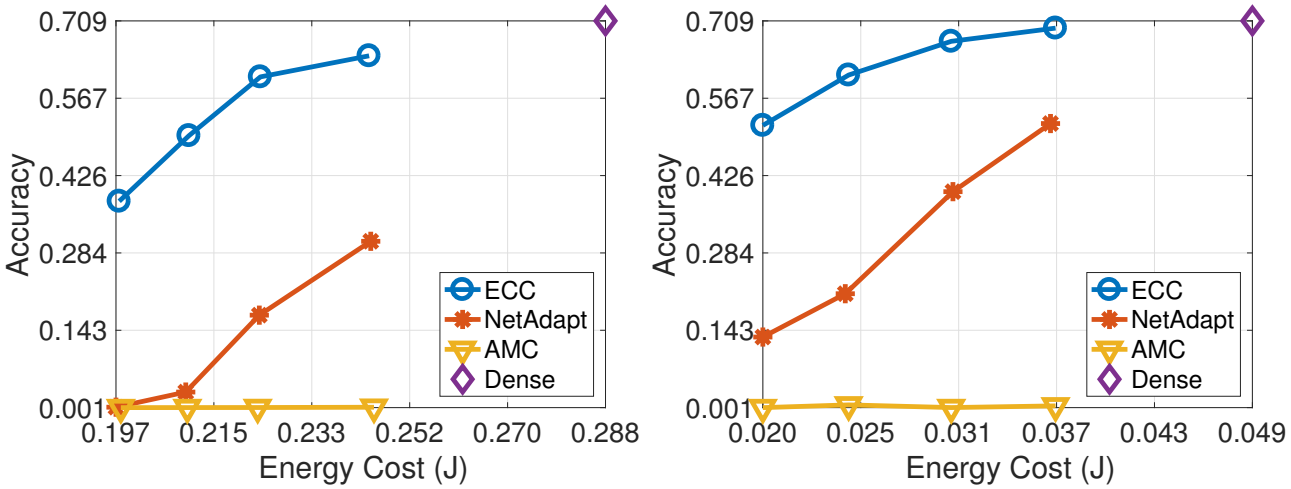
Supplementary Material

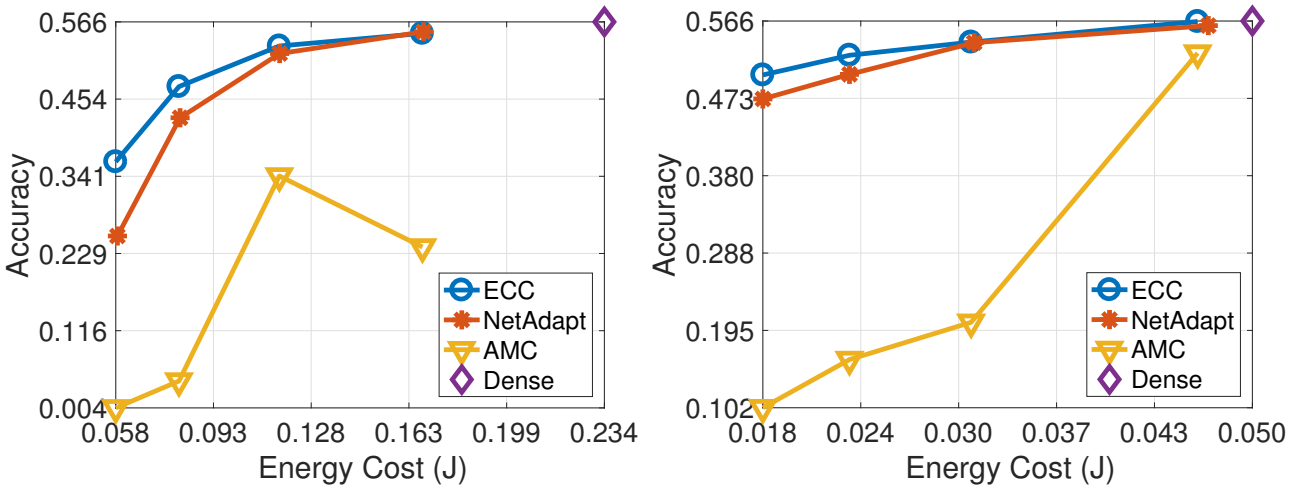
More Experiment Results

ImageNet classification accuracy before fine-tuning The accuracy gains are significant before fine-tuning as shown in Figure 7 and Figure 8. For instance, on TX2 under the same 0.037 J energy budget, ECC achieves 17.5% higher accuracy on MobileNet compared to NetAdapt. AMC has lower accuracies since it does not update DNN parameters in the RL searching phase. Overall, we find that ECC is insensitive to additional fine-tuning while both NetAdapt and AMC require extensive fine-tuning to improve accuracy. This is because ECC, through its constrained optimization process, inherently performs compression and fine-tuning simultaneously.



(a) GTX 1080 Ti. (b) Jetson TX2.

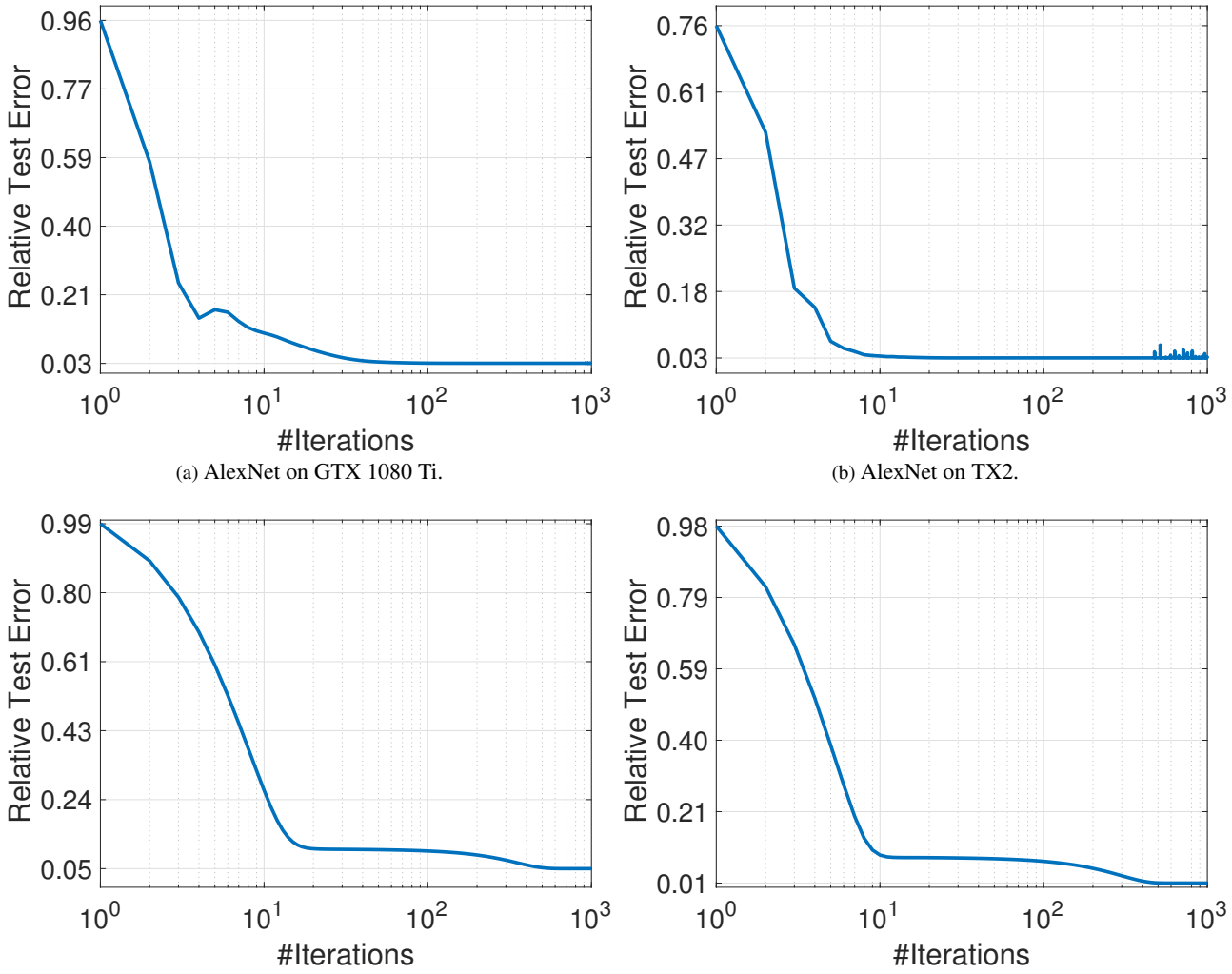
Figure 7: Top-1 accuracy of image classification on MobileNet@ImageNet **before** fine-tuning.



(a) GTX 1080 Ti. (b) Jetson TX2.

Figure 8: Top-1 accuracy of image classification on AlexNet@ImageNet **before** fine-tuning.

Energy model prediction errors on other networks Figures 9 and 10 show the results of the energy prediction models as in Section 4.2.



(a) AlexNet on GTX 1080 Ti.

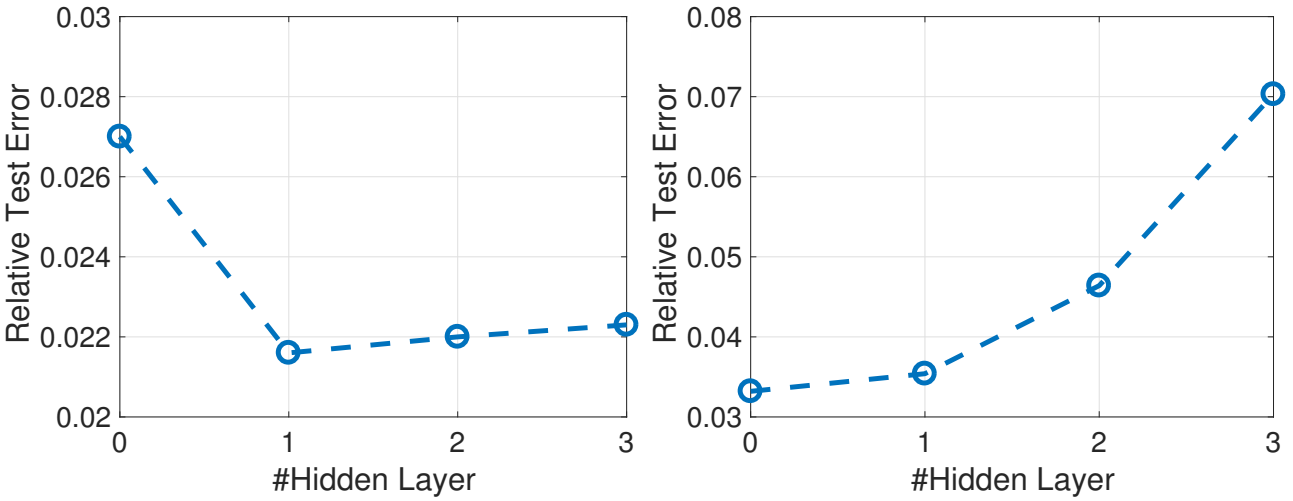
(b) AlexNet on TX2.

(c) ERFNet on GTX 1080 Ti.

(d) ERFNet on TX2.

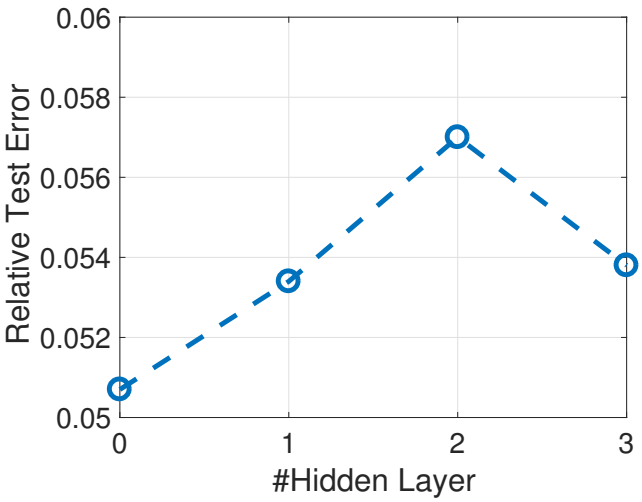
Figure 9: Relative test error of energy prediction using the proposed bilinear model.

Layer (inverse) sparsity on other networks Figures 11 shows the layer (inverse) sparsity of the complementary compressed models of Figure 6. For MobileNet on GTX 1080 Ti, the lower bound of $s^{(u)}$ is set to be $0.35c^{(u)}$.

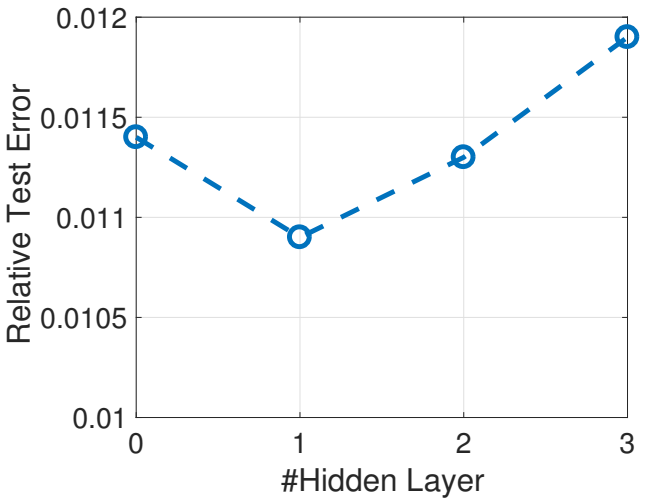


(a) AlexNet on GTX 1080 Ti.

(b) AlexNet on TX2.

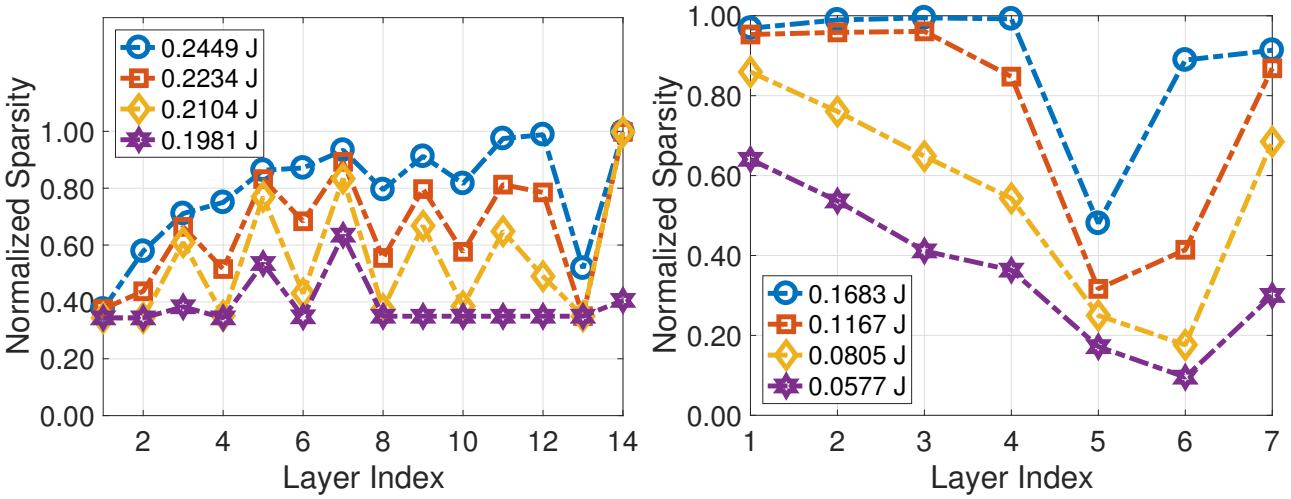


(c) ERFNet on GTX 1080 Ti.



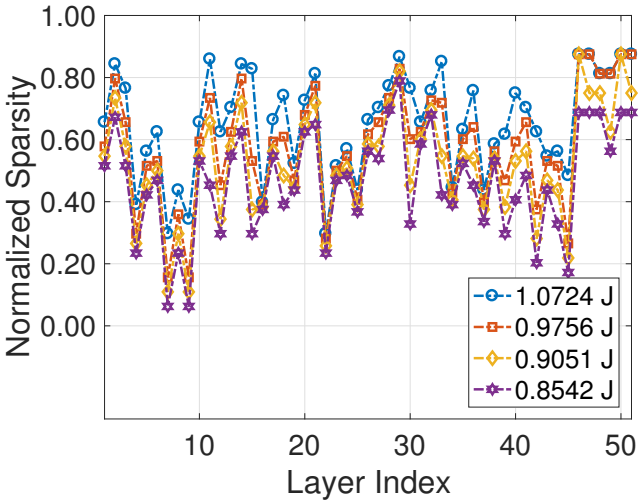
(d) ERFNet on TX2.

Figure 10: Relative test error of energy prediction using an MLP model with different hidden layers.

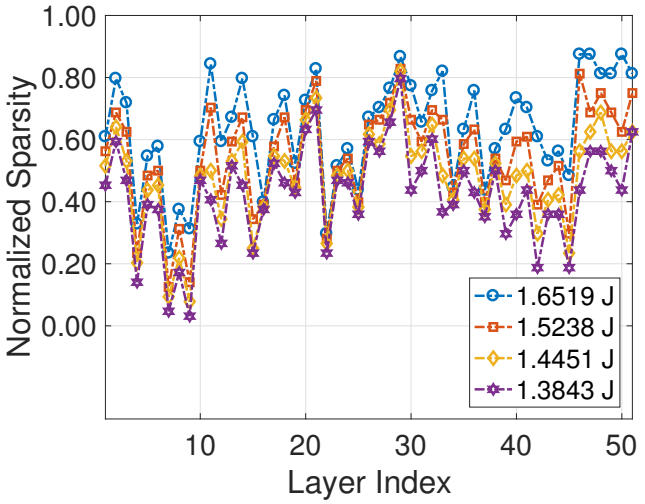


(a) MobileNet on GTX 1080 Ti.

(b) AlexNet on GTX 1080 Ti.



(c) ERFNet on TX2.



(d) ERFNet on GTX 1080 Ti.

Figure 11: Layer (inverse) sparsity after compressing.