We thank the reviewers for their attention and constructive feedback that will improve the quality of our work.

Reviewer 1: We acknowledge the need for clearer referencing to the SM and will improve this. Also, the study of Novak et al. 2019 on the role of locality is indeed relevant; we will discuss it in the revised version.

We agree that studying the effect of further training methods and architectures would be very interesting. Such exploration is certainly an exciting direction for new research. One practical obstacle for different architectures lies in the implementation of the mapping for each layer. Our work presents a proof of concept that will hopefully trigger several investigations along the lines proposed by the reviewer.

One step we took in this direction, motivated by reviewers' suggestions, is to replicate the same experiment with Adam instead of fixed learning rate SGD. This additional result, which we will add in the revised version, confirms the phenomena shown in Figure 1 of the SM.

Further subjects we would like to study in a new paper: the effects of locality and weight sharing; embedding of CNNs into other CNNs with bigger filters; explore the effects of data augmentation on the off-diagonal blocks; scale the experiments to near SOTA models on CIFAR10; and more...

Reviewer 2: We did our best to motivate the fact that relaxing the constraints at the right point is a promising training technique, by showing the performance improvement in two simple setups on CIFAR-10 and CIFAR-100. To strengthen our claims, we are working on implementing a softer constraint relaxation by mapping to locally-connected space rather the fully-connected space. This approach has stronger practical benefits as the increase in the model size is much less than the eFCN embedding.

In the revised version we will make clearer the potential implications of our method to practitioners, and its foreseeable extensions. We agree that this emphasis will be complementary to the study of the effects of architectural bias in and of itself.

Reviewer 3: We acknowledge that the "VanillaCNN" model used on CIFAR-10 is rather small, nevertheless its generalization performance is almost the same as simplified AlexNet on CIFAR-10. Our strategy has been to present the results for the VanillaCNN in the main text and then validate them for more realistic setups in the SM. The reviewer may have overlooked that the experiment with AlexNet is performed on CIFAR-100 (not CIFAR-10 as for the VanillaCNN), which is why the test accuracy is $\sim 40\%$. Furthermore, in our experience, the best tuned FCN on CIFAR-10 hardly beats $\sim 60\%$. The main reason we present the VanillaCNN in the main text is that it is practically unfeasible to perform the Hessian analysis on AlexNet (with our computational constraints). In the revised version we will clarify better the link between the results presented in the main text and the SM.

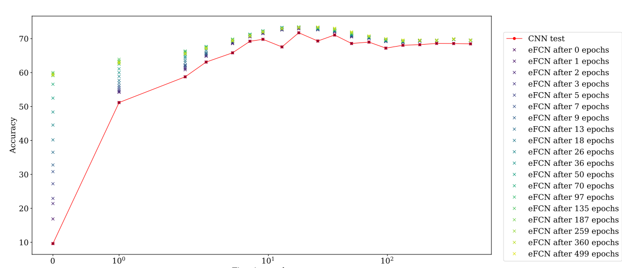

Figure 1: Same as Fig.2a of the SM, but with Adam optimizer used for both the CNN and the eFCNs (initial learning rate of 0.001).

We understand the very valid concern of the reviewer about the learning rate scheduling, and had actually considered this question, although we do not discuss it in the main text. If our understanding is correct, the reviewer suspects that the fact that the learning rate is divided by 10 upon switching gives the eFCN an unfair advantage over the CNN which keeps a constant learning rate. However, learning rates are intrinsically related to model sizes, and considering how different the CNN and the eFCN are in size, it would be a tricky (yet interesting) question to define what would be a "fair" learning rate to use for the eFCNs. Therefore we solely chose learning rates of 0.1 for the CNNs and 0.01 for the eFCNs so that the corresponding models would all converge in a reasonable and comparable timescale of the order of 100 epochs. Motivated by the reviewer's comment, we repeated our numerical experiments using the adaptive Adam optimizer to circumvent this question. Fig. 1 shows the results obtained in that case and confirms our conclusions. We will include this additional finding in the revision.

In Ba & Caruana 2013, the shallow (and fully-connected) model is trained by regressing a previously trained deep (and convolutional) model, whereas in our case the fully-connected models benefit from the architectural bias only through the initial stages of training. We thank the reviewer for the pointer. We will add it to our discussion of papers related to model compression.