We thank all reviewers for their feedback. We have fixed minor issues/typos raised by reviewers.

R1 Results on NLP and Audio Tasks In the limited rebuttal window, we were able to achieve promising results on MLPs using β-LASSO on NLP and Audio tasks: Test accuracy on UrbanSound8K (audio classification): a)MLP (β-LASSO): 61.5% b)MLP (SGD+weight decay): 54.2%. Test accuracy on AG News (text classification): a)MLP (β-LASSO): 91.2% b)MLP (SGD+weight decay): 88.1% (SOTA is XLNet (Yang et. al. 2019) with 95.5%). We will add more results on other domains in the final version. Since this was the only major concern raised by the reviewer, we respectfully ask them to consider accepting this paper.

R2 Introducing S-FC, S-CONV We first want to clarify that as the title suggests, this paper is not about the "impact of sparsity in training neural networks". It is about understanding the inductive bias of convolutions and moving towards learning them from data. One of the contributions of the paper is to showcase the success of an algorithm that encourages sparsity in learning local connectivity. We hope that reviewer would judge the paper according to the main motivation and contributions of the paper. In section 2.1 we discuss in details why introducing S-FC and S-CONV is necessary in order to do a *controlled study* on the inductive bias of convolutions.

Choice of benchmarks Again as we explain in the introduction, the goal of the paper is to understand the inductive bias of the convolutions and use that to bridge the gap between MLPs and convnets (which ends up learning local connections). As a consequence, we compare our method against best reported results on training MLPs. Are are not aware of any other work that has a competitive result for training MLPs on these datasets.

Comparison to LASSO and other algorithms/regularizers that encourage sparsity We have reported the best known results on training MLPs using *any* algorithm/regularizer. Since our algorithm is a variant of LASSO (which we use loosely to refer to L1-regularization), we have also compared it in Table 2 ($\beta = 0$ corresponds to LASSO). To strengthen the results, we will implement a few other algorithms/regularizers that encourage sparsity and test them in the final version. Thanks for the suggestions!

This research domain has moved on from toy experiments... Again we want to emphasize that this paper is not about presenting yet another algorithm to sparsify networks but it is a study of inductive bias of convolutions and is presenting sparsity as a way to learn local connectivity from data instead of directly incorporating it in the architecture. Training MLPs on ImageNet is very challenging and we are not aware of any such successful attempts.

R3 Theorem 1 MDL theorem is well-known and we only state it formally in Theorem 1 for completion. We have clearly added ([26, Theorem 7.7]) right before the statement which is the convention to refer to a theorem taken from another work. As you pointed $d(h)$ in equation 1 should be replaced with $|d(h)|$ to refer to the length of $d(h)$. We will improve the discussions around Theorem 1 in the final version.

#param in local and convolutional counterparts Local and convolution counterparts have the same number of weights but the convolutional counterpart has less parameters since it uses weights sharing to assign a group of weights to the same parameter.

Double Descent It has been shown that double descent can be mitigated with regularization (Nakkiran et al. 2020). We do not observe double descent phenomenon in our experiments (not even the epoch-wise double descent). We will add discussions to address this in the paper.

R4 It is known to research community that models with fewer parameters generalize better While this is very much motivated by VC-dimension and other generalization theory, it is well-documented in deep learning literature that models with more parameters often generalize better (the opposite of what theory suggests). Here, what we meant is to show that this deviation from theory only happens in over-parameterized regime and therefore to compare two model families (based on their generalization), we suggest that one can compare them in under-parameterized regime and the model family that performs better on under-parametrized regime will also perform better in over-parameterized regime. We are not aware of any prior work that points to this observation about such scaling behavior of model families in deep learning. We will add more context to this discussion in the paper to make it more clear.

Transferability of the search operation Thanks for these suggestions. We are indeed thinking of these as future work. However, we want to emphasize on the significance of the current result. Even without going deep, we have improved the SOTA on training MLPs by 10% on CIFAR10 dataset (to 84.5 using a simple algorithm. Being able to learn convolutions takes many such papers and we are hoping to inspire the community to work on these directions. Considering these contributions and with clarifications we have provided, we hope that the reviewer would consider accepting the paper.