Author Response to Reviews

Thank you for your time in reading the paper and the positive feedback! Below are responses for each reviewer.

Response to Reviewer (id) 1

Thank you for your detailed reading of the paper and positive feedback!

Similarity of representations from random init across random runs: The greater variation in representational similarity for CBR-LargeT and CBR-LargeW likely arises due to the randomness of the initialization and the training process across different runs, which can affect similarity scores (see e.g. SVCCA, (Raghu et al, NeurIPS 2017)). It is therefore striking that for all the other models, the representations cluster so clearly into orange and blue in the plots from the paper – suggesting significant differences between representations learned from pretraining vs random initialization. Even for CBR-LargeT and CBR-LargeW, the mean similarity across the different runs is larger for the blue dots compared to the orange dots. With even more runs, we expect this distinction will be even clearer.

Referenced paper: Thank you for the reference to the paper on developing CNNs for tuberculosis screening and visualization, which we will add to the related work. We think it provides an interesting example of a non-standard architecture in the medical domain, but there are significant differences between it and our work. In particular, the paper states that 'the use of pretraining is outside the scope of the paper' – and does not compare the effect of pretrained weights vs random initialization. This is the core of the transfer learning question, and a central part of our paper. Additionally, the focus of the referenced paper is a specific medical application – tuberculosis – while our paper studies properties of pretraining and variation in architectures from representational viewpoints and across datasets, studying the effect of model overparametrization, the absence of traditional features such as Gabor filters, transfer in limited data regimes and convergence speed and pure scaling benefits of transfer.

Response to Reviewer (id) 3

Thank you for the positive feedback, and for the Figure 3 reference typo comment (which is indeed what we meant) and microaneurysm color – we will edit these in the updated version!

CCA similarity in Figures 2, 3: The fact that CCA similarity in Figure 2 is lower than Figure 3 is likely due to two reasons. Firstly, multiple papers [1] Convergent Learning, (Li, Yosinski, Clune, Lipson, Hopcroft, ICLR 2016), [2] SVCCA, (Raghu et al, NeurIPS 2017), [3] Towards understanding learning representations (Wang, Hu, Gu, Wu, Hu, He, Hopcroft, NeurIPS 2018) have shown that deep networks trained on the same task are much less similar in the middle layers than at the start. This is even more the case when the networks are initialized differently (as suggested by Figure 2), and so when computing an average similarity over multiple layers, we reduce the overall CCA similarity score compared to just the first layer. Secondly, the CCA similarity score has some mild variation in scale across layers, due to different layers having different numbers of neurons (further discussed in [4] Insights on Representational Similarity, (Morcos, Raghu, Bengio, NeurIPS 2018), and [5] Similarity of Neural Network Representations Revisited (Kornblith, Norouzi, Lee, Hinton, ICML 2019). Therefore, a comparison of the CCA similarity scores across Figures 2, 3 might be harder to interpret directly (while the comparisons within Figures 2, 3 have a clear correspondence.)

Page 7 lines 197-199: Our intuition of larger networks moving less comes from Figure 3 and the filter visualizations in Figure 4. This phenomenon is consistent with related effects we have found since the submission in papers that we reference in the updated version of the paper: [6] Neural Tangent Kernel, (Jacot et al), [7] Are all layers created equal? (Zhang et al). In the submission, we computed the Euclidean distance between parameters at the start and end of training to further study this effect. However, Euclidean distance is sensitive to the total number of parameters and the resulting differences in initialization scalings, which we tried to address with a heuristic of dividing by the initialization norm $||w_0||$. But particularly in high dimensions, this remains a coarse measure of distance traveled, and while it was enough to show a difference between e.g. Resnet and the CBRs, it is not fine-grained enough to capture differences within the CBR family. In future work we hope to derive a more fine-grained distance metric.

Response to Reviewer (id) 5

Thank you for the detailed review and positive feedback!

Small Dataset Size: We performed experiments on this important question with results shown in Table 3 in the Appendix of the submission. We simulated the small data regime by training our models on a small subset of the full training set but testing on the full test set. Our results highlight the following conclusions: for larger, overparametrized models e.g. Resnet50, there is a larger gap between transfer learning and random initialization for the smallest training dataset sizes (e.g. 5k training points). However, for smaller models, *both transfer and random initialization perform the same*, even for the 5k datapoints setting. This suggests that some of the gains from transfer learning in the small data regime are due to the overparametrization of the transfer-defined model architecture, rather than the reuse of pretrained features to prevent overfitting on the small dataset. In the updated version, we will add these conclusions to the main text.