Transfusion: Understanding Transfer Learning for Medical Imaging

Maithra Raghu[*]
Cornell University and Google Brain
maithrar@gmail.com

Chiyuan Zhang[*]
Google Brain
chiyuan@google.com

Jon Kleinberg[†]
Cornell University
kleinber@cs.cornell.edu

Samy Bengio[†]
Google Brain
bengio@google.com

Abstract

Transfer learning from natural image datasets, particularly ImageNet, using standard large models and corresponding pretrained weights has become a de-facto method for deep learning applications to medical imaging. However, there are fundamental differences in data sizes, features and task specifications between natural image classification and the target medical tasks, and there is little understanding of the effects of transfer. In this paper, we explore properties of transfer learning for medical imaging. A performance evaluation on two large scale medical imaging tasks shows that surprisingly, transfer offers little benefit to performance, and simple, lightweight models can perform comparably to ImageNet architectures. Investigating the learned representations and features, we find that some of the differences from transfer learning are due to the over-parametrization of standard models rather than sophisticated feature reuse. We isolate where useful feature reuse occurs, and outline the implications for more efficient model exploration. We also explore feature independent benefits of transfer arising from weight scalings.

1 Introduction

With the growth of deep learning, transfer learning has become integral to many applications — especially in medical imaging, where the present standard is to take an existing architecture designed for natural image datasets such as ImageNet, together with corresponding pretrained weights (e.g. ResNet [10], Inception [27]), and then fine-tune the model on the medical imaging data.

This basic formula has seen almost universal adoption across many different medical specialties. Two prominent lines of research have used this methodology for applications in radiology, training architectures like ResNet, DenseNet on chest x-rays [31, 24] and ophthalmology, training Inception-v3, ResNet on retinal fundus images [2, 9, 23, 4]. The research on ophthalmology has also culminated in FDA approval [28], and full clinical deployment [29]. Other applications include performing early detection of Alzheimer's Disease [5], identifying skin cancer from dermatologist level photographs [6], and even determining human embryo quality for IVF procedures [15].

Despite the immense popularity of transfer learning in medical imaging, there has been little work studying its precise effects, even as recent work on transfer learning in the *natural image* setting

[*]Equal Contribution.
[†]Equal Contribution.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

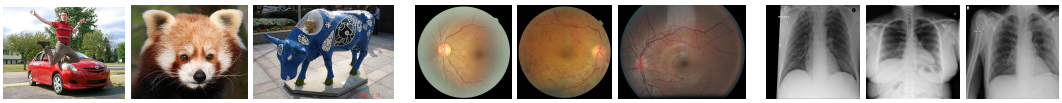


Figure 1: Example images from the *ImageNet*, the *retinal fundus photographs*, and the CheXpert datasets, respectively. The fundus photographs and chest x-rays have much higher resolution than the ImageNet images, and are classified by looking for small local variations in tissue.

[11, 16, 20, 12, 7] has challenged many commonly held beliefs. For example in [11], it is shown that transfer (even between similar tasks) does not necessarily result in performance improvements, while [16] illustrates that pretrained features may be less general than previously thought.

In the medical imaging setting, many such open questions remain. As described above, transfer learning is typically performed by taking a standard ImageNet architecture along with its pretrained weights, and then fine-tuning on the target task. However, ImageNet classification and medical image diagnosis have considerable differences.

First, many medical imaging tasks start with a large image of a bodily region of interest and use variations in local textures to identify pathologies. For example, in retinal fundus images, small red 'dots' are an indication of microaneurysms and diabetic retinopathy [1], and in chest x-rays local white opaque patches are signs of consolidation and pneumonia. This is in contrast to natural image datasets like ImageNet, where there is often a clear global subject of the image (Fig. 1). There is thus an open question of how much ImageNet feature reuse is helpful for medical images.

Additionally, most datasets have larger images (to facilitate the search for local variations), but with many fewer images than ImageNet, which has roughly one million images. By contrast medical datasets range from several thousand images [15] to a couple hundred thousand [9, 24].

Finally, medical tasks often have significantly fewer classes (5 classes for Diabetic Retinopathy diagnosis [9], 5 – 14 chest pathologies from x-rays [24]) than the standard ImageNet classification setup of 1000 classes. As standard ImageNet architectures have a large number of parameters concentrated at the higher layers for precisely this reason, the design of these models is likely to be suboptimal for the medical setting.

In this paper, we perform a fine-grained study on transfer learning for medical images. Our main contributions are:

[1] We evaluate the performance of standard architectures for natural images such as ImageNet, as well as a family of non-standard but smaller and simpler models, on two large scale medical imaging tasks, for which transfer learning is currently the norm. We find that (i) in all of these cases, transfer does not significantly help performance (ii) smaller, simpler convolutional architectures perform comparably to standard ImageNet models (iii) ImageNet performance is not predictive of medical performance. These conclusions also hold in the very small data regime.

[2] Given the comparable performance, we investigate whether using pretrained weights leads to different learned representations, by using (SV)CCA [22] to directly analyze the hidden representations. We find that pretraining does affect the hidden representations, but there is a confounding issue of model size, where the large, standard ImageNet models do not change significantly through the fine-tuning process, as evidenced through surprising correlations between representational similarity at initialization and after convergence.

[3] Using further analysis and weight transfusion experiments, where we partially reuse pretrained weights, we isolate locations where meaningful feature reuse does occur, and explore hybrid approaches to transfer learning where a subset of pretrained weights are used, and other parts of the network are redesigned and made more lightweight.

[4] We show there are also *feature-independent* benefits to pretraining — reusing only the *scaling* of the pretrained weights but not the features can itself lead to large gains in convergence speed.

2 Datasets

Our primary dataset, the Retina data, consists of retinal *fundus photographs* [9], large 587 × 587 images of the back of the eye. These images are used to diagnose a variety of eye diseases including

Diabetic Retinopathy (DR) [3]. DR is graded on a five-class scale of increasing severity [1]. Grades 3 and up are *referable DR* (requiring immediate specialist attention), while grades 1 and 2 correspond to *non-referable DR*. As in prior work [9, 2] we evaluate via AUC-ROC on identifying referable DR.

We also study a second medical imaging dataset, CheXpert [14], which consists of chest x-ray images (resized to 224 × 224), which can be used to diagnose 5 different thoracic pathologies: atelectasis, cardiomegaly, consolidation, edema and pleural effusion. We evaluate our models on the AUC of diagnosing each of these pathologies. Figure 1 shows some example images from both datasets and ImageNet, demonstrating drastic differences in visual features among those datasets.

3 Models and Performance Evaluation of Transfer Learning

To lay the groundwork for our study, we select multiple neural network architectures and evaluate their performance when (1) training from random initialization and (2) doing transfer learning from ImageNet. We train both standard, high performing ImageNet architectures that have been popular for transfer learning, as well as a family of significantly smaller convolutional neural networks, which achieve comparable performance on the medical tasks.

As far as we are aware, there has been little work studying the effects of transfer learning from ImageNet on smaller, non-standard ImageNet architectures. (For example, [21] studies a different model, but does not evaluate the effect of transfer learning.) This line of investigation is especially important in the medical setting, where large, computationally expensive models might significantly impede mobile and on-device applications. Furthermore, in standard ImageNet models, most of the parameters are concentrated at the top, to perform the 1000-class classification. However, medical diagnosis often has considerably fewer classes – both the retinal fundus images and chest x-rays have just 5 classes – likely meaning that ImageNet models are highly overparametrized.

We find that across both datasets and all models, transfer learning does not significantly affect performance. Additionally, the family of smaller lightweight convolutional networks performs comparably to standard ImageNet models, despite having significantly worse accuracy on ImageNet– the ImageNet task is not necessarily a good indication of success on medical datasets. Finally, we observe that these conclusions also hold in the setting of very limited data.

3.1 Description of Models

For the standard ImageNet architectures, we evaluate ResNet50 [11] and Inception-v3 [27], which have both been used extensively in medical transfer learning applications [2, 9, 31]. We also design a family of simple, smaller convolutional architectures. The basic building block for this family is the popular sequence of a (2d) convolution, followed by batch normalization [13] and a relu activation. Each architecture has four to five repetitions of this basic layer. We call this model family CBR. Depending on the choice of the convolutional filter size (fixed for the entire architecture), the number of channels and layers, we get a family of architectures with size ranging from a third of the standard ImageNet model size (CBR-LargeT, CBR-LargeW) to one twentieth the size (CBR-Tiny). Full architecture details are in the Appendix.

3.2 Results

We evaluate three repetitions of the different models and initializations (random initialization vs pretrained weights) on the two medical tasks, with the result shown in Tables 1, 2. There are two possibilities for repetitions of transfer learning: we can have a fixed set of pretrained weights and multiple training runs from that initialization, or for each repetition, first train from scratch on ImageNet and then fine-tune on the medical task. We opt for evaluating the former, as that is the standard method used in practice. For all models except for Inceptionv3, we first train on ImageNet to get the pretrained weights. For Inceptionv3, we used the pretrained weights provided by [26].

Table 1 shows the model performances on the Retina data (AUC of identifying moderate Diabetic Retinopathy (DR), described in Section 2), along with ImageNet top 5 accuracy. Firstly, we see that transfer learning has minimal effect on performance, not helping the smaller CBR architectures at all, and only providing a fraction of a percent gain for Resnet and Inception. Next, we see that despite the significantly lower performance of the CBR architectures on ImageNet, they perform very comparably to Resnet and Inception on the Retina task. These same conclusions are seen

Dataset	Model Architecture	Random Init	Transfer	Parameters	ImageNet Top5
Retina	Resnet-50	96.4% ± 0.05	96.7% ± 0.04	23570408	92.% ± 0.06
Retina	Inception-v3	96.6% ± 0.13	96.7% ± 0.05	22881424	93.9%
Retina	CBR-LargeT	96.2% ± 0.04	96.2% ± 0.04	8532480	77.5% ± 0.03
Retina	CBR-LargeW	95.8% ± 0.04	95.8% ± 0.05	8432128	75.1% ± 0.3
Retina	CBR-Small	95.7% ± 0.04	95.8% ± 0.01	2108672	67.6% ± 0.3
Retina	CBR-Tiny	95.8% ± 0.03	95.8% ± 0.01	1076480	73.5% ± 0.05

Table 1: **Transfer learning and random initialization perform comparably across both standard ImageNet architectures and simple, lightweight CNNs for AUCs from diagnosing moderate DR. Both sets of models also have similar AUCs, despite significant differences in size and complexity.** Model performance on DR diagnosis is also not closely correlated with ImageNet performance, with the small models performing poorly on ImageNet but very comparably on the medical task.

Model Architecture	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
Resnet-50	79.52±0.31	75.23±0.35	85.49±1.32	88.34±1.17	88.70±0.13
Resnet-50 (trans)	79.76±0.47	74.93±1.41	84.42±0.65	88.89±1.66	88.07±1.23
CBR-LargeT	81.52±0.25	74.83±1.66	88.12±0.25	87.97±1.40	88.37±0.01
CBR-LargeT (trans)	80.89±1.68	76.84±0.87	86.15±0.71	89.03±0.74	88.44±0.84
CBR-LargeW	79.79±0.79	74.63±0.69	86.71±1.45	84.80±0.77	86.53±0.54
CBR-LargeW (trans)	80.70±0.31	77.23±0.84	86.87±0.33	89.57±0.34	87.29±0.69
CBR-Small	80.43±0.72	74.36±1.06	88.07±0.60	86.20±1.35	86.14±1.78
CBR-Small (trans)	80.18±0.85	75.24±1.43	86.48±1.13	89.09±1.04	87.88±1.01
CBR-Tiny	80.81±0.55	75.17±0.73	85.31±0.82	84.87±1.13	85.56±0.89
CBR-Tiny (trans)	80.02±1.06	75.74±0.71	84.28±0.82	89.81±1.08	87.69±0.75

Table 2: **Transfer learning provides mixed performance gains on chest x-rays.** Performances (AUC%) of diagnosing different pathologies on the CheXpert dataset. Again we see that transfer learning does not help significantly, and much smaller models performing comparably.

on the chest x-ray results, Table 2. Here we show the performance AUC for the five different pathologies (Section 2). We again observe mixed gains from transfer learning. For Atelectasis, Cardiomegaly and Consolidation, transfer learning performs slightly worse, but helps with Edema and Pleural Effusion.

3.3 The Very Small Data Regime

We conducted additional experiments to study the effect of transfer learning in the very small data regime. Most medical datasets are significantly smaller than ImageNet, which is also the case for our two datasets. However, our datasets still have around two hundred thousand examples, and other settings many only have a few thousand. To study the effects in this very small data regime, we trained models on only 5000 datapoints on the Retina dataset, and examined the effect of transfer learning. The results, in Table 3, suggest that while transfer learning has a bigger effect

Model	Rand Init	Pretrained
Resnet50	92.2%	94.6%
CBR-LargeT	93.6%	93.9%
CBR-LargeW	93.6%	93.7%

Table 3: **Benefits of transfer learning in the small data regime are largely due to architecture size.** AUCs when training on the Retina task with only 5000 datapoints. We see a bigger gap between random initialization and transfer learning for Resnet (a large model), but not for the smaller CBR models.

with very small amounts of data, there is a confounding effect of model size – transfer primarily helps the large models (which are designed to be trained with a million examples) and smaller models again show little difference between transfer and random initialization.

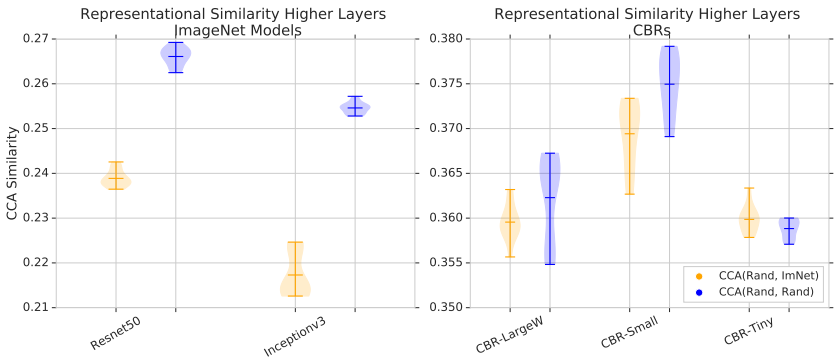


Figure 2: **Pretrained weights give rise to different hidden representations than training from random initialization for large models.** We compute CCA similarity scores between representations learned using pretrained weights and those from random initialization. We do this for the top two layers (or stages for Resnet, Inception) and average the scores, plotting the results in orange. In blue is a baseline similarity score, for representations trained from different random initializations. We see that representations learned from random initialization are more similar to each other than those learned from pretrained weights for larger models, with less of a distinction for smaller models.

4 Representational Analysis of the Effects of Transfer

In Section 3 we saw that transfer learning and training from random initialization result in very similar performance across different neural architectures and tasks. This gives rise to some natural questions about the effect of transfer learning on the kinds of *representations* learned by the neural networks. Most fundamentally, does transfer learning in fact result in any representational differences compared to training from random initialization? Or are the effects of the initialization lost? Does feature reuse take place, and if so, where exactly? In this section, we provide some answers to these basic questions. Our approach directly analyzes and compares the hidden representations learned by different populations of neural networks, using (SV)CCA [22, 19], revealing an important dependence on *model size*, and differences in behavior between lower and higher layers. These insights, combined with results Section 5 suggest new, hybrid approaches to transfer learning.

Quantitatively Studying Hidden Representations with (SV)CCA To understand how pretraining affects the features and representations learned by the models, we would like to (quantitatively) study the learned intermediate functions (latent layers). Analyzing latent representations is challenging due to their complexity and the lack of any simple mapping to inputs, outputs or other layers. A recent tool that effectively overcomes these challenges is (Singular Vector) Canonical Correlation Analysis, (SV)CCA [22, 19], which has been used to study latent representations through training, across different models, alternate training objectives, and other properties [22, 19, 25, 18, 8, 17, 30]. Rather than working directly with the model parameters or neurons, CCA works with *neuron activation vectors* — the ordered collection of outputs of the neuron on a sequence of inputs. Given the activation vectors for two sets of neurons (say, corresponding to distinct layers), CCA seeks linear combinations of each that are as correlated as possible. We adapt existing CCA methods to prevent the size of the activation sets from overwhelming the computation in large models (details in Appendix C), and apply them to compare the latent representations of corresponding hidden layers of different pairs of neural networks, giving a CCA similarity score of the learned intermediate functions.

Transfer Learning and Random Initialization Learn Different Representations Our first experiment uses CCA to compare the similarity of the hidden representations learned when training from pretrained weights to those learned when training from random initialization. We use the representations learned at the top two layers (for CBRs) or stages (for Resnet, Inception) before the output layer, averaging their similarity scores. As a baseline to compare to, we also look at CCA similarity scores for the same representations when training from random initialization with two different seeds (different initializations and gradient updates). The results are shown in Figure 2. For larger models (Resnet, Inception), there is a clear difference between representations, with the similarity of representations between training from random initialization and pretrained

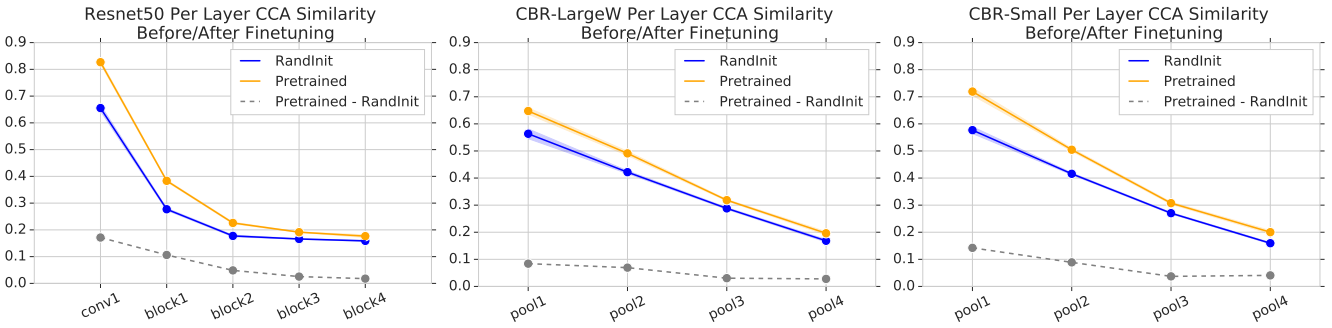


Figure 3: **Per-layer CCA similarities before and after training on medical task.** For all models, we see that the lowest layers are most similar to their initializations, and this is especially evident for Resnet50 (a large model). We also see that feature reuse is mostly restricted to the bottom two layers (stages for Resnet) — the only place where similarity with initialization is significantly higher for pretrained weights (grey dotted lines shows the difference in similarity scores between pretrained and random initialization).

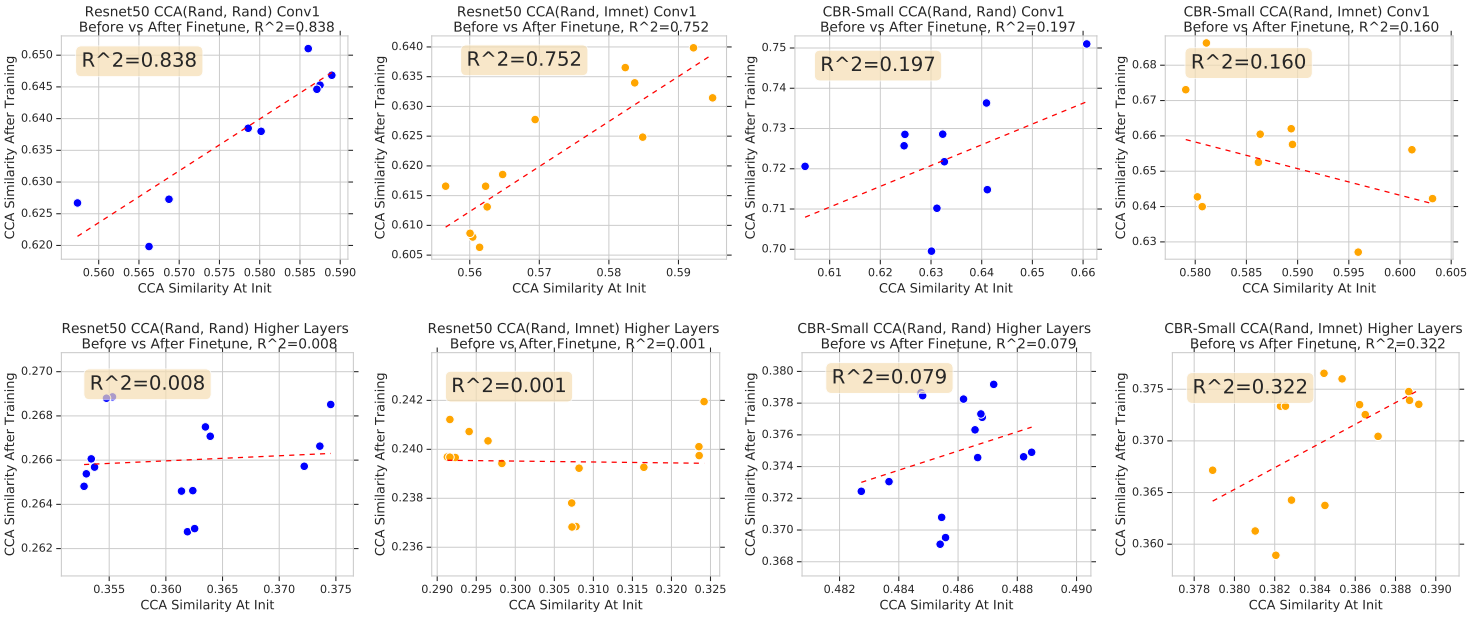
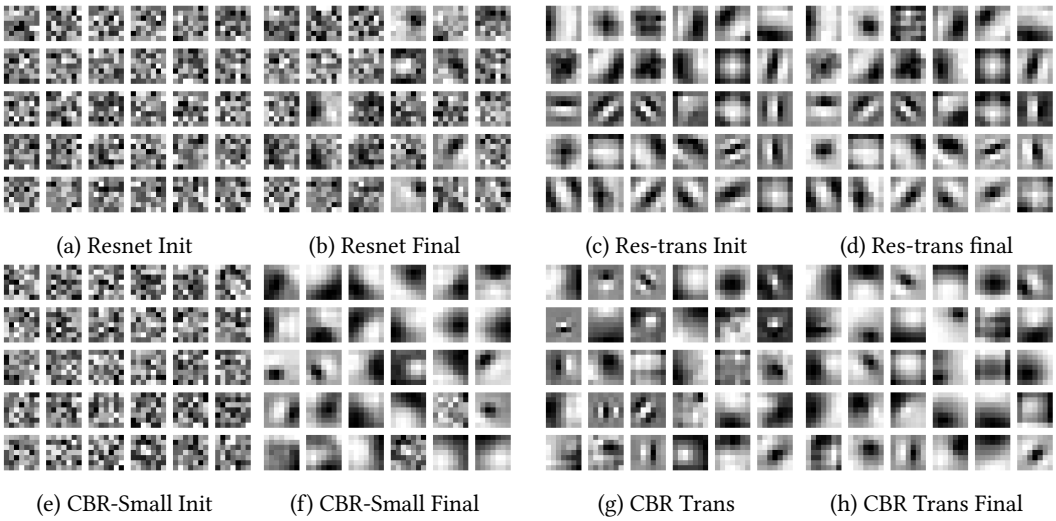


Figure 4: **Large models move less through training at lower layers: similarity at initialization is highly correlated with similarity at convergence for large models.** We plot CCA similarity of Resnet (conv1) initialized randomly and with pretrained weights at (i) initialization, against (ii) CCA similarity of the converged representations (top row second from left.) We also do this for two different random initializations (top row, left). In *both* cases (even for random initialization), we see a surprising, strong correlation between similarity at initialization and similarity after convergence (R^2 = 0.75, 0.84). This is not the case for the smaller CBR-Small model, illustrating the overparametrization of Resnet for the task. Higher must likely change much more for good task performance.

weights (orange) noticeably lower than representations learned independently from different random initializations (blue). However for smaller models (CBRs), the functions learned are more similar.

Larger Models Change Less Through Training The reasons underlying this difference between larger and smaller models becomes apparent as we further study the hidden representations of all the layers. We find that *larger models change much less during training*, especially in the lowest layers. This is true *even when they are randomly initialized*, ruling out feature reuse as the sole cause, and implying their overparametrization for the task. This is in line with other recent findings [33].



(a) Resnet Init (b) Resnet Final (c) Res-trans Init (d) Res-trans final

(e) CBR-Small Init (f) CBR-Small Final (g) CBR Trans (h) CBR Trans Final

Figure 5: **Visualization of conv1 filters shows the remains of initialization after training in Resnet, and the lack of and erasing of Gabor filters in CBR-Small.** We visualize the filters before and after training from random initialization and pretrained weights for Resnet (top row) and CBR-Small (bottom row). Comparing the similarity of (e) to (f) and (g) to (h) shows the limited movement of Resnet through training, while CBR-Small changes much more. We see that CBR does not learn Gabor filters when trained from scratch (f), and also erases some of the pretrained Gabors (compare (g) to (h).)

In Figure 3, we look at per-layer representational similarity before/after finetuning, which shows that the lowest layer in Resnet (a large model), is significantly more similar to its initialization than in the smaller models. This plot also suggests that any serious feature reuse is restricted to the lowest couple of layers, which is where similarity before/after training is clearly higher for pretrained weights vs random initialization. In Figure 4, we plot the CCA similarity scores between representations using pretrained weights and random initialization *at initialization* vs after training, for the lowest layer (conv1) as well as higher layers, for Resnet and CBR-Small. Large models changing less through training is evidenced by a surprising correlation between the CCA similarities for Resnet conv1, which is not true for higher layers or the smaller CBR-Small model.

Filter Visualizations and the Absence of Gabors As a final study of how pretraining affects the model representations, we visualize some of the filters of conv1 for Resnet and CBR-Small (both 7x7 kernels), before and after training on the Retina task. The filters are shown in Figure 5, with visualizations for chest x-rays in the Appendix. These add evidence to the aformentioned observation: the Resnet filters change much less than those of CBR-Small. In contrast, CBR-Small moves more from its initialization, and has more similar learned filters in random and pretrained initialization. Interestingly, CBR-Small does not appear to learn Gabor filters when trained from scratch (bottom row second column). Comparing the third and fourth columns of the bottom row, we see that CBR-Small even erases some of the Gabor filters that it is initialized with in the pretrained weights.

5 Convergence: Feature Independent Benefits and Weight Transfusion

In this section, we investigate the effects of transfer learning on convergence speed, finding that: (i) surprisingly, transfer offers *feature independent* benefits to convergence simply through better weight scaling (ii) using pretrained weights from the lowest two layers/stages has the biggest effect on convergence — further supporting the findings in the previous section that any meaningful feature reuse is concentrated in these lowest two layers (Figure 3.) These results suggest some hybrid approaches to transfer learning, where only a subset of the pretrained weights (lowest layers) are used, with a lightweight redesign to the top of the network, and even using entirely *synthetic* features, such as synthetic Gabor filters (Appendix F.3). We show these hybrid approaches capture most of the benefits of transfer and enable greater flexibility in its application.

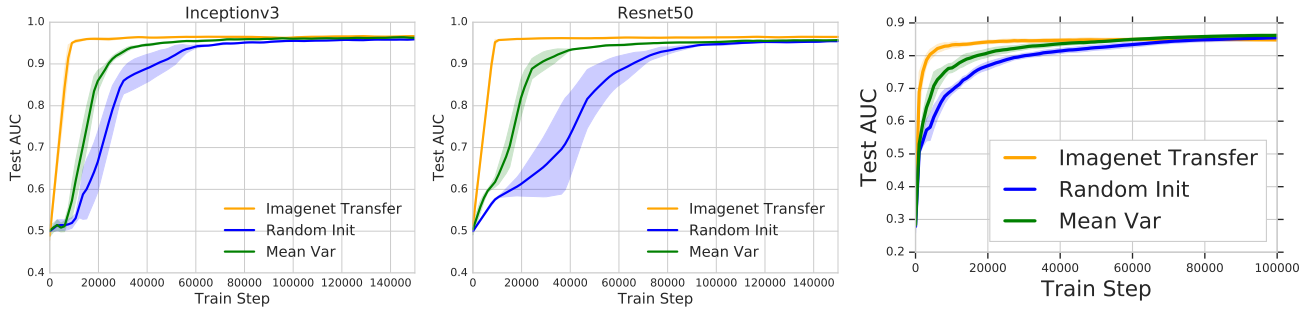


Figure 6: **Using only the scaling of the pretrained weights (Mean Var Init) helps with convergence speed**. The figures compare the standard transfer learning and the *Mean Var* initialization scheme to training from scratch. On both the RETINA data (a-b) and the CHEXPERT data (c) (with Resnet50 on the *Consolidation* disease), we see convergence speedups.

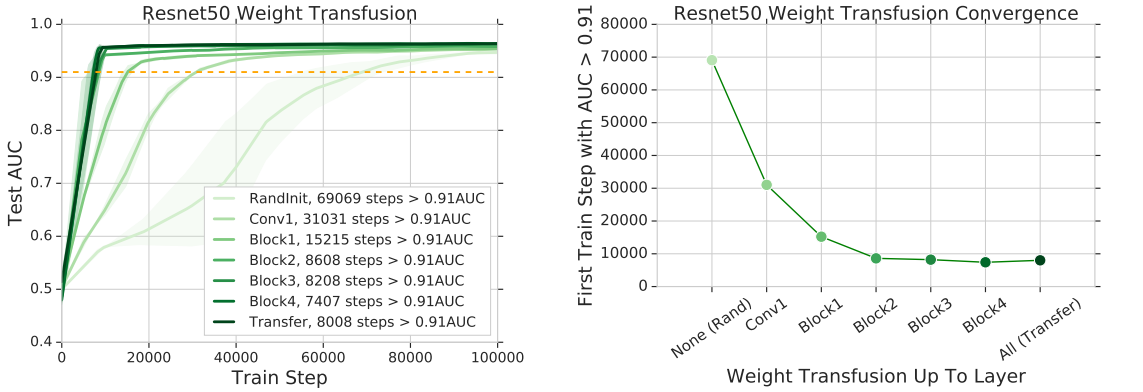


Figure 7: **Reusing a subset of the pretrained weights (weight transfusion), further supports only the lowest couple of layers performing meaningful feature reuse.** We initialize a Resnet with a contiguous subset of the layers using pretrained weights (weight transfusion), and the rest randomly, and train on the RETINA task. On the left, we show the convergene plots when transfusing up to conv1 (just one layer), up to block 1 (conv1 and all the layers in block1), etc up to full transfer. On the right, we plot the number of train steps taken to reach 91% AUC for different numbers of transfused weights. Consistent with findings in Section 4, we observe that reusing the lowest layers leads to the greatest gain in convergence speed. Perhaps surprisingly, just reusing conv1 gives the greatest marginal convergence speedup, even though transfusing weights for a block means several new layers are using pretrained weights.

Feature Independent Benefits of Transfer: Weight Scalings We consistently observe that using pretrained weights results in faster convergence. One explanation for this speedup is that there is significant feature reuse. However, the results of Section 4 illustrate that there are many confounding factors, such as model size, and feature reuse is likely limited to the lowest layers. We thus tested to see whether there were *feature independent* benefits of the pretrained weights, such as *better scaling*. In particular, we initialized a *iid weights* from $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, where $\tilde{\mu}$ and $\tilde{\sigma}^2$ are the mean and variance of $\tilde{W}$, the pretrained weights. Doing this for each layer separately inherits the scaling of the pretrained weights, but destroys all of the features. We called this the *Mean Var* init, and found that it significantly helps speed up convergence (Figure 6.) Several additional experiments studying batch normalization, weight sampling, etc are in the Appendix.

Weight Transfusions and Feature Reuse We next study whether the results suggested by Section 4, that meaningful feature reuse is restricted to the lowest two layers/stages of the network is supported by the effect on convergence speed. We do this via a *weight transfusion* experiment, transfering a contiguous set of some of the pretrained weights, randomly initializing the rest of the network, and training on the medical task. Plotting the training curves and steps taken to reach a threshold AUC in Figure 7 indeed shows that using pretrained weights for lowest few layers has the biggest training speedup. Interestingly, just using pretrained weights for conv1 for Resnet results

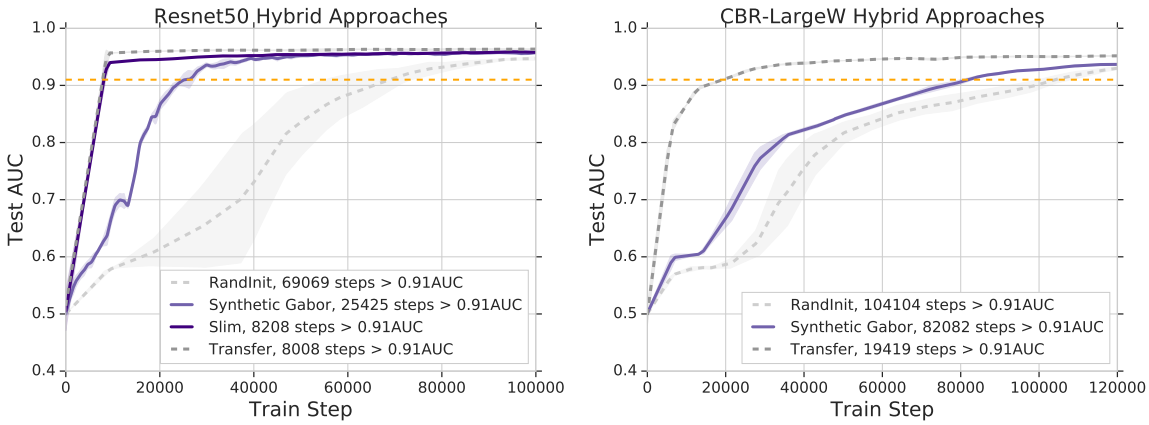


Figure 8: **Hybrid approaches to transfer learning: reusing a subset of the weights and slimming the remainder of the network, and using synthetic Gabors for conv1.** For Resnet, we look at the effect of reusing pretrained weights up to Block2, and slimming the remainder of the network (halving the number of channels), randomly initializing those layers, and training end to end. This matches performance and convergence of full transfer learning. We also look at initializing conv1 with *synthetic Gabor filters* (so *no* use of pretrained weights), and the rest of the network randomly, which performs equivalently to reusing conv1 pretrained weights. This result generalizes to different architectures, e.g. CBR-LargeW on the right.

in the largest gain, despite transfusion for a Resnet block meaning multiple layers are now reusing pretrained weights.

Takeaways: Hybrid Approaches to Transfer Learning The transfusion results suggest some hybrid, more flexible approaches to transfer learning. Firstly, for larger models such as Resnet, we could consider reusing pretrained weights up to e.g. Block2, redesiging the top of the network (which has the bulk of the parameters) to be more lightweight, initializing these layers randomly, and training this new Slim model end to end. Seeing the disproportionate importance of conv1, we might also look at the effect of initializing conv1 with *synthetic Gabor filters* (see Appendix F.3 for details) and the rest of the network randomly. In Figure 8 we illustrate these hybrid approaches. Slimming the top of the network in this way offers the same convergence and performance as transfer learning, and using synthetic Gabors for conv1 has *the same effect* as pretrained weights for conv1. These variants highlight many new, rich and flexible ways to use transfer learning.

6 Conclusion

In this paper, we have investigated many central questions on transfer learning for medical imaging applications. Having benchmarked both standard ImageNet architectures and non-standard lightweight models (itself an underexplored question) on two large scale medical tasks, we find that transfer learning offers limited performance gains and much smaller architectures can perform comparably to the standard ImageNet models. Our exploration of representational similarity and feature reuse reveals surprising correlations between similarities at initialization and after training for standard ImageNet models, providing evidence of their overparametrization for the task. We also find that meaningful feature reuse is concentrated at the lowest layers and explore more flexible, hybrid approaches to transfer suggested by these results, finding that such approaches maintain all the benefits of transfer and open up rich new possibilities. We also demonstrate feature-independent benefits of transfer learning for better weight scaling and convergence speedups.

References

[1] AAO. *International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table.* American Academy of Ophthalmology, 2002.

[2] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly

available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13):5200–5206, 2016.

[3] Hasseb Ahsan. Diabetic retinopathy – biomolecules and multiple pathophysiology. *Diabetes and Metabolic Syndrome: Clincal Research and Review*, pages 51–54, 2015.

[4] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

[5] Yiming Ding, Jae Ho Sohn, Michael G Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W Jenkins, Dmytro Lituiev, Timothy P Copeland, Mariam S Aboian, Carina Mari Aparici, et al. A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 290(2):456–464, 2018.

[6] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[8] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

[9] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip Q Nelson, Jessica Mega, and Dale Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018.

[12] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[15] Pegah Khosravi, Ehsan Kazemi, Qiansheng Zhan, Marco Toschi, Jonas E Malmsten, Cristina Hickman, Marcos Meseguer, Zev Rosenwaks, Olivier Elemento, Nikica Zaninovic, et al. Robust automated assessment of human blastocyst quality using deep learning. *bioRxiv*, page 394882, 2018.

[16] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*, 2018.

[17] Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*, 2019.

[18] Martin Magill, Faisal Qureshi, and Hendrick de Haan. Neural networks trained to solve differential equations learn general representations. In *Advances in Neural Information Processing Systems*, pages 4075–4085, 2018.

[19] Ari S Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *arXiv preprint arXiv:1806.05759*, 2018.

[20] Jiquan Ngiam, Daiyi Peng, Vijay Vasudevan, Simon Kornblith, Quoc V Le, and Ruoming Pang. Domain adaptive transfer learning with specialist models. *arXiv preprint arXiv:1811.07056*, 2018.

[21] F Pasa, V Golkov, F Pfeiffer, D Cremers, and D Pfeiffer. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific reports*, 9(1):6268, 2019.

[22] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017.

[23] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction with applications to healthcare. *arXiv preprint arXiv:1807.01771*, 2018.

[24] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, arXiv:1711.05225, 2017.

[25] Naomi Saphra and Adam Lopez. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*, 2018.

[26] Tensorflow Slim. Tensorflow slim inception-v3. `https://github.com/tensorflow/models/tree/master/research/slim`, 2017.

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[28] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44, 2019.

[29] Amber A Van Der Heijden, Michael D Abramoff, Frank Verbraak, Manon V van Hecke, Albert Liem, and Giel Nijpels. Validation of automated screening for referable diabetic retinopathy with the idx-dr device in the hoorn diabetes care system. *Acta ophthalmologica*, 96(1):63–68, 2018.

[30] Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. *arXiv preprint arXiv:1909.01380*, 2019.

[31] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE, 2017.

[32] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[33] Chiyuan Zhang, Samy Bengio, and Yoram Singer. Are all layers created equal? *arXiv preprint arXiv:1902.01996*, 2019.

Appendix to "Transfusion: Understanding Transfer Learning for Medical Imaging"

A Details on Datasets, Models and Hyperparameters

The Retina dataset consisted of around 250k training images, and 70k test images. The train test split was done by patient id (as is standard for medical datasets) to ensure no accidental similarity between the train/test dataset. The chest x-ray dataset is open sourced and available from [14], which has all of the details. Briefly, they have 223k training images and binary indicator for multiple diseases assiciated with each image extracted automatically from the meta data. The standard ImageNet (ILSVRC 2012) dataset was also used to pretrain models.

For dataset preprocessing we used mild random cropping, as well as standard normalization by the mean and standard deviation for ImageNet. We augmented the data with hue and contrast augmentations. For the Retina data, we used random horizontal and vertical flips, and for the chest x-ray data, we did not do random flip. We did not do model specific hyperparameter tuning on each target data, and used fixed standard hyperparameters.

For experiments on the Retina data, we trained the standard ImageNet models, Resnet50 and Inception-v3, by replacing the final 1000 class ImageNet classification head with a five class head for DR diagnosis, or five classes for the five different chest x-ray diseases. We use the sigmoid activation at the top instead of the multiclass softmax activation, and the train the models in the multi-label binary classification framework.

The CBR family of small convolutional neural networks consists of multiple conv2d-batchnorm-relu layers followed by a maxpool. Each maxpool has spatial window (3x3) and stride (2x2). For each CBR architecture, there is one filter size for all the convolutions (which all have stride 1). Below, conv-n denotes a 2d convolutionl with n output channels.

- **CBR-LargeT**(all) has 7x7 conv filters: (conv32-bn-relu) maxpool (conv64-bn-relu) maxpool (conv128-bn-relu) maxpool (conv256-bn-relu) maxpool (conv512-bn-relu) global avgpool, classification
- **CBR-LargeW**(ide) has 7x7 conv filters: (conv64-bn-relu) maxpool (conv128-bn-relu) maxpool (conv256-bn-relu) maxpool (conv512-bn-relu) maxpool, global avgpool, classification.
- **CBR-Small** has 7x7 conv filters: (conv32-bn-relu) maxpool (conv64-bn-relu) maxpool (conv128-bn-relu) maxpool (conv256-bn-relu) maxpool global avgpool, classification
- **CBR-Tiny** has 5x5 conv filters: (conv64-bn-relu) maxpool (conv128-bn-relu) maxpool (conv256-bn-relu) maxpool (conv512-bn-relu) maxpool, global avgpool, classification.

The models on Retina are trained on 587×587 images, with learning rate 0.001 and a batch size of 8 (for memory considerations.) The Adam optimizer is used. The models on the chest x-ray are trained on 224×224 images, with a batch size of 32, and vanilla SGD with momentum (coefficient 0.9). The learning rate scheduling is inherited from the ImageNet training pipeline, which warms up from 0 to $0.1 \times \frac{32}{256}$ in 5 epochs, and then decay with a factor of 10 on epoch 30, 60, and 90, respectively.

B Additional Dataset Size Results

Complementing the data varying experiments in the main text, we additional experiments on varying the amount of training data, fidning that for around 50k datapoints, we return to only seeing a fractional improvement of transfer learning. Future work could study how hybrid approaches perform when less data is available.

C CCA Details

For full details on the implementation of CCA, we reference prior work [22, 19], as well as the open sourced code (the source of our implementation): https://github.com/google/svcca

Model	Init Method	5k	10k	50k	100k
Resnet50	ImageNet Pretrained	94.6%	94.8%	95.7%	96.0
Resnet50	Random Init	92.2%	93.3%	95.3%	95.9%
CBR-LargeT	Random Init	93.6%	-	-	-
CBR-LargeT	Pretrained	93.9%	-	-	-
CBR-LargeW	Random Init	93.6%	-	-	-
CBR-LargeW	Pretrained	93.7%	-	-	-
Resnet50	Conv1 Pretrained	92.9%	-	-	-
Resnet50	Mean Var Init	-	94.4%	95.5%	95.8%

Table 4: **Additional performance results when varying initialization and the dataset size on the Retina task.** For Resnet50, we show performances when training on very small amounts of data. We see that even finetuning (with early stopping) on 5k datapoints beats the results from performing fixed feature extraction, Figure 12, suggesting finetuning should always be preferred. For 5k, 10k datapoints, we see a larger gap between transfer learning and random init (closed by 50k datapoints) but this is likely due to the enormous size of the model (typically trained on 1 million datapoints) compared to the dataset size. This is supported by evaluating the effect of transfer on CBR-LargeT and CBR-LargeW, where transfer again does not help much. (These are one third the size of Resnet50, and we expect the gains of transfer to be even more minimal for CBR-Small and CBR-Tiny.) We also show results for using the MeanVar init, and see some gains in performance for the very small data setting. We also see a small gain on 5k datapoints when just reusing the conv1 weights for Resnet50.

One challenge we face when implementing CCA is the large size of the convolutional activations. These activations have shape (n, h, w, c), where n is the number of datapoints, c the number of channels, and h, w the spatial dimensions. These values all vary significantly across the network, e.g. conv1 has shape $(n, 294, 294, 64)$, while activations at the end of block 3 have shape $(n, 19, 19, 1024)$. Because CCA is sensitive to both the number of datapoints n (actually hwn for convolutional layers) and the number of neurons – c for large convolutional layers – there is large variations in scaling across different layers in the model. To address this, we do the following: let L and L' be the layers we want to compare, with shape (height, width, channels), (h_L, w_L, c_L). We apply CCA as follows:

- Pick p, the total number of image patches to compute activation vectors and CCA over, and d, the maximum number of neuron activation vectors to correlate with
- Pick the number of datapoints n so that $nh_L w_L = p$.
- Sample d of the c_L channels, and apply CCA to the resulting d x $nh_L w_L$ activation matrices.
- Repeat over samples of d and n.

This works much better than prior approaches of averaging over all of the spatial dimensions [19], or flattening across all of the neurons [22] (too computationally expensive in this setting.)

D Additional Results from Representation Analysis

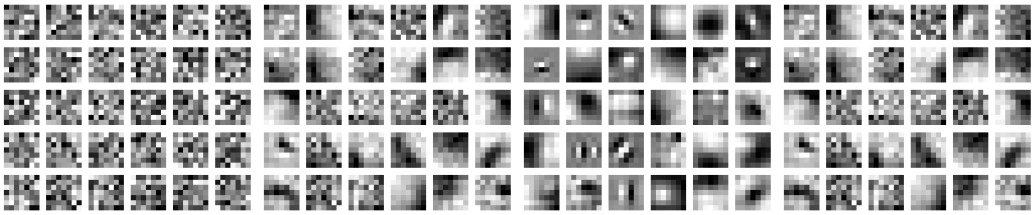
Here, we include some additional results studying the representations of these models. We perform more representational similarity comparisons between networks trained from (the same) pretrained weights (as is standard), but different random seeds. We do this for Resnet50 (a large model) and CBR-Small (a small model), and Table 5 includes these results as well as similarity comparisons for networks trained with different random seeds and *different* random initializations as a baseline. The comparisons across layers and models is slightly involved, but as we detail below, the evidence further supports the conclusions in the main text:

- *Larger models change less through training.* Comparing CCA similarity scores across models is a little challenging, due to different scalings, so we look at the difference in CCA similarity between two models trained with pretrained weights, and two models trained from random initialization, for Resnet50 and CBR-Small. Comparing this value for Conv1 (in Resnet50) to Pool1 (in CBR-Small), we see that pretraining results in much more similar representations compared to random initialization in the large model over the small model.

Description	Conv1	Block1	Block2	Block3	Block4
Resnet50 CCA(ImNet1, ImNet2)	0.865	0.559	0.421	0.343	0.313
Resnet50 CCA(Rand1, Rand2)	0.647	0.369	0.277	0.256	0.276
Resnet50 Diff	**0.218**	**0.191**	**0.144**	**0.086**	**0.037**

Description	Pool1	Pool2	Pool3	Pool4
CBR-Small CCA(ImNet1, ImNet2)	0.825	0.709	0.477	0.395
CBR-Small CCA(Rand1, Rand2)	0.723	0.541	0.401	0.349
CBR-Small Diff	**0.102**	**0.168**	**0.076**	**0.046**

Table 5: **Representational comparisons between trained ImageNet models with different seeds highlight the variation of behavior in higher and lower layers, and differences between larger and smaller models.** We compute CCA similarity between representations at different layers when training from different random seeds with (i) (the same) pretrained weights (ii) different random inits, for Resnet and CBR-Small. The results support the conclusions of the main text. For Resnet50, in the lowest layers such as Conv1 and Block1, we see that representations learned when using (the same) pretrained weights are much more similar to each other (diff 0.2 in CCA score) than representations learned from different random initializations. This ~ 0.2 difference is also much higher than (somewhat) corresponding differences in CBR-Small, for Pool1, Pool2. Actually, as Resnet50 is much deeper, the large difference in Block1 is very striking. (Block 1 alone contains much more layers than all of CBR-Small.) By Block3 and Block4 however, the CCA similarity difference between pretrained representations and those from random initialization is much smaller, and slightly lower than the differences for Pool3, Pool4 in CBR-Small, suggesting that pretrained weights are not having much of a difference on the kinds of functions learned. For CBR-Small, we also see that pretrained weights result in larger differences between the representations in the lower layers, but these become much smaller in the higher layers. We also observe that representations in CBR-Small trained from random initialization (especially in the lower layers e.g. Pool1) are more similar to each other than in Resnet50, suggesting things move more.



(a) rand init (b) final (rand init) (c) transfer init (d) final (transfer init)

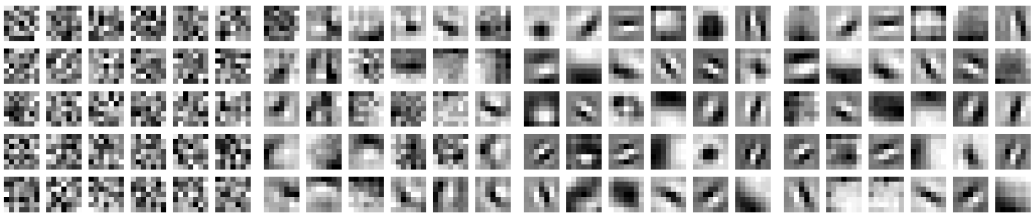
Figure 9: **First layer filters of CBR-Small on the CheXpert data.** (a) and (c) show the randomly initialized filters and filters initialized from a model (the same architecture) pre-trained on ImageNet. (b) and (d) shows the final converged filters from the two different initializations, respectively.

- *The effect of pretraining is mostly limited to the lowest layers* For higher layers, the CCA similarities between representations using pretrained weights and those trained from random initializations are closer, and the difference between CBR-Small and Resnet-50 is non-existent, suggesting that the effects of pretraining mostly affect the lowest layers across models, with finetuning changing representations at the top.

Figure 9 and Figure 10 compare the first layer filters between transfer learning and training from random initialization on the CheXpert data for the CBR-Small and Resnet-50 architectures, respectively. Those results complement Figure 5 in the main text.

E The Fixed Feature Extraction Setting

To complete the picture, we also study the fixed feature extractor setting. While the most popular methodology for transfer learning is to initialize from pretrained weights and fine-tune (train) the entire network, an alternative is to initialize all layers up to layer L with pretrained weights. These are then treated as a fixed feature extractor, with only layers $L + 1$ onwards, being trained. There are two variants of this fixed feature extractor experiment: **[1]** Initialize all layers with pretrained



(a) rand init (b) final (rand init) (c) transfer init (d) final (transfer init)

Figure 10: **First layer filters of Resnet-50 on the CheXpert data.** (a) and (c) show the randomly initialized filters and filters initialized from a model (the same architecture) pre-trained on ImageNet. (b) and (d) shows the final converged filters from the two different initializations, respectively.

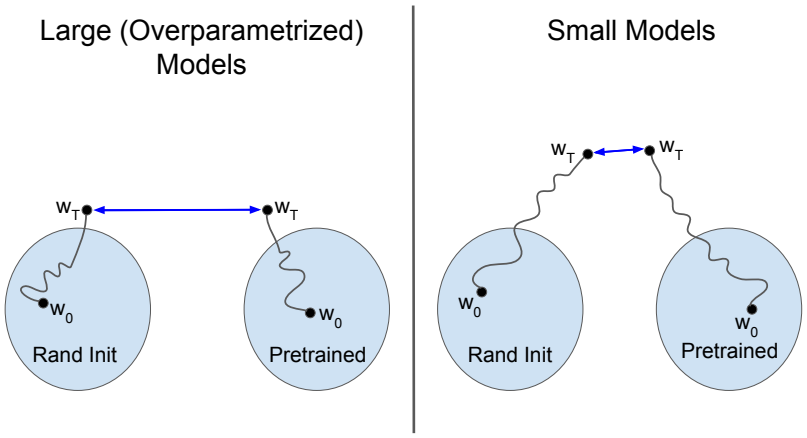


Figure 11: **Larger models move less through training than smaller networks.** A schematic diagram of our intuition for optimization for larger and smaller models.

weights and only train layer $L + 1$ onwards. **[2]** Initialize only up to layer L with pretrained weights, and layer $L + 1$ onwards randomly; then train only layers $L + 1$ onwards.

We implement both of these versions across different models trained on the Retina task in Figure 12, and CheXpert in Figure 13, including a baseline of using random features – initializing the network randomly, freezing up to layer L, and training layer $L + 1$ onwards. For the Retina task, we see that the pretrained ImageNet features perform significantly better than the random features baseline, but this gap is significantly closer on the chest x-rays.

More surprisingly however, there is little difference in performance between initializing all layers with pretrained weights and only up to layer L with pretrained weights. This latter experiment has also been studied in [32], where they found that re-initializing caused drops in performance due to *co-adaptation*, where neurons in different layers have evolved together in a way that is not easily discoverable through retraining. This analysis was done for highly similar tasks (different subsets of ImageNet), and we hypothesise that in our setting, the significant changes of the higher layers (Figures 3, 4) means that the correct adaptation is naturally learned through training.

F Additional Results on Feature Independent Benefits and Weight Transfusions

Figure 14 visualizes the first layer filters from various initialization schemes. As shown in the main text, the *Mean Var* initialization could converge much faster than the baseline random initialization due to better parameter scaling transferred from the pre-trained weights. Figure 15 shows more

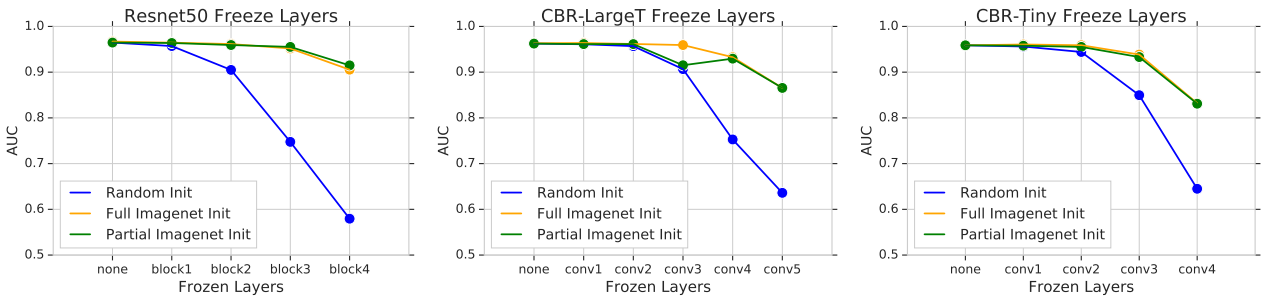


Figure 12: **ImageNet features perform well as fixed feature extractors on the Retina task, and are robust to coadaptation performance drops.** We initialize (i) the full architecture with ImageNet weights (yellow) (ii) up to layer L with ImageNet weights, and the rest randomly. In both, we keep up to layer L fixed, and only train layers $L+1$ onwards. We compare to a random features baseline, initializing randomly and training layer $L+1$ onwards (blue). ImageNet features perform much better as fixed feature extractors than the random baseline (though this gap is much closer for the CheXpert dataset, Appendix Figure 13.) Interestingly, there is no performance drop due to the *coadaptation* issue [32], with partial ImageNet initialization performing equally to initialzing with all of the ImageNet weights.

results on Retina with various architectures. We find that on smaller models, the effectiveness of the *Mean Var* initialization is less very pronounced, likely due to them being much shallower.

Figure 16 shows all the five diseases on the CheXpert data for Resnet-50. Except for Cardiomegaly, we see benefits of the *Mean Var* initialization scheme on convergence speed in all other diseases.

F.1 Batch Normalization Layers

Batch normalization layers Ioffe and Szegedy [13] are an essential building block for most modern network architectures with visual inputs. However, these layers have a slightly different structure that requires more careful consideration when performing the Mean Var init. Letting x be a batch of activations, batch norm computes

$$\gamma\left(\frac{(x-\mu_B)}{\sigma_B+\epsilon}\right)+\beta$$

Here, γ, β are learnable scale, shift parameters, and μ_B, σ_B are an accumulated running mean and variance over the train dataset. Thus, in transfer learning, μ_B, σ_B start off as the mean/variance of the ImageNet data activations, unlikely to match the medical image statistics. Therefore, for the Mean Var Init, we initialized all of the batch norm parameters to the identity: $\gamma, \sigma_B = 1, \beta, \mu_B = 0$. We call this the *BN Identity Init.* Two alternatives are *BN ImageNet Mean Var*, resampling the values of all batch norm parameters according to the ImageNet means and variances, and *BN ImageNet Transfer*, copying over the batch norm parameters from ImageNet. We compare these three methods in Figure 17, with non-batchnorm layers initialized according to the Mean Var Init. Broadly, they perform similarly, with *BN Identity Init* (used by default in other Mean Var related experiments) performing slightly better. We observe that *BN ImageNet Transfer*, where the ImageNet batchnorm parameters are transferred directly to the medical images, performs the worst.

F.2 Mean Var Init vs Using Knowledge of the Full Empirical ImageNet Weight Distribution

In Figure 14, we see that while the Mean Var Init might have the same mean and variance as the ImageNet weight distribution, the two distributions themselves are quite different from each other. We examined the convergence speed of initializing with the Mean Var Init vs initializing using knowledge of the entire empirical distribution of the ImageNet weights.

In particular, we looked at (1) *Sampling Init:* each weight is drawn iid from the full empirical distribution of ImageNet weights (2) *Shuffled Init:* random shuffle of the pretrained ImageNet weights to form a new initialization. (Note this is exactly sampling from the empirical distribution

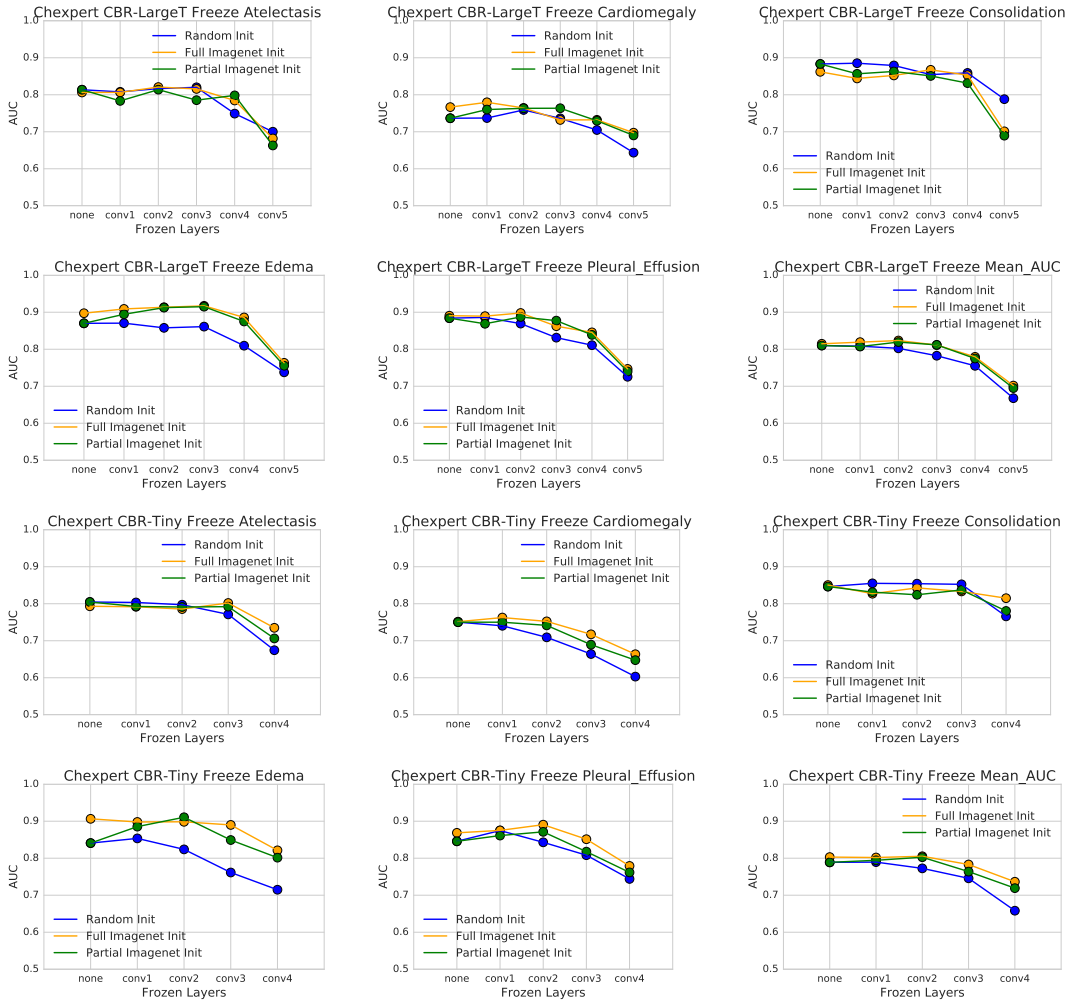


Figure 13: **Experiments on freezing lower layers of CBR-LargeT and a CBR-Tiny model on the CheXpert data.** After random or transfer initialization, we keep up to layer L fixed, and only train layers $L+1$ onwards. ImageNet features perform better as fixed feature extractors than the random baseline for most diseases, but the gap is much closer than for the Retina data, Figure 12. We again see that there is no significant performance drop due to coadaptation challenges.

without replacement.) The results are illustrated in Figure 18. Interestingly, Mean Var is very similar in convergence speed to both of these alternatives. This would suggest that further improvements in convergence speed might have to come from also modelling correlations between weights.

F.3 Synthetic Gabor Filters

We test mathematically synthetic Gabor filters in place of learned Gabor filters on ImageNet for its benefits in speeding up the convergence when used as initialization in the first layer of neural networks. The Gabor filters are generated with the `skimage` package, using the following code snippet.

```
from skimage.filters import gabor_kernel
from skimage.transform import resize
import numpy as np

def gen_gabors(n_angles=16, sigmas=[2], freqs = [0.08, 0.16, 0.25, 0.32],
               kernel_resize = 10, kernel_crop = 7):
```

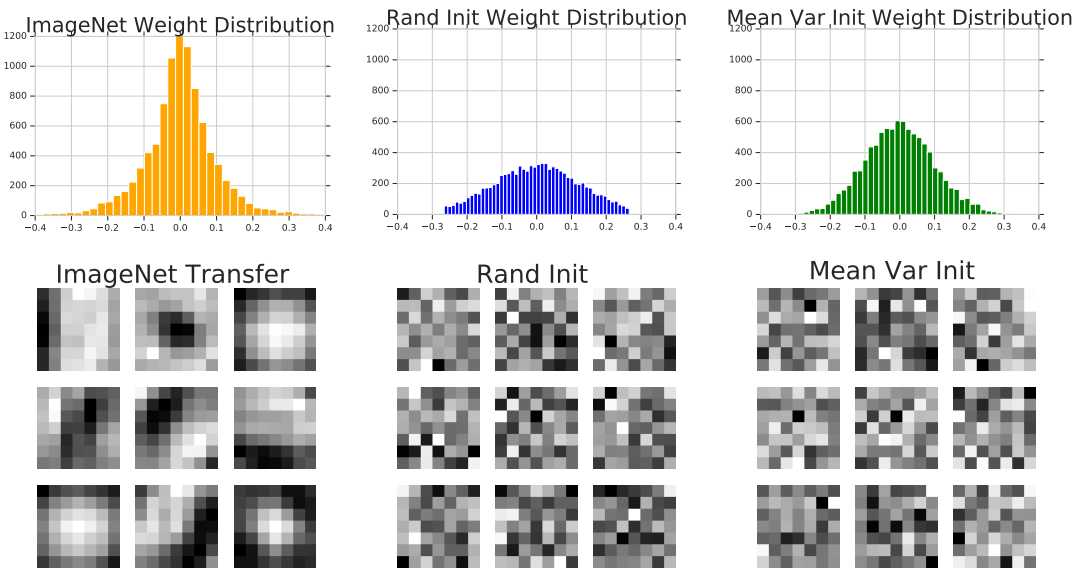


Figure 14: **Distribution and filter visualization of weights initialized according to pretrained ImageNet weights, Random Init, and Mean Var Init**. The top row is a histogram of the weight values of the the first layer of the network (Conv 1) when initialized with these three different schemes. The bottom row shows some of the filters corresponding to the different initializations. Only the ImageNet Init filters have pretrained (Gabor-like) structure, as Rand Init and Mean Var weights are iid.

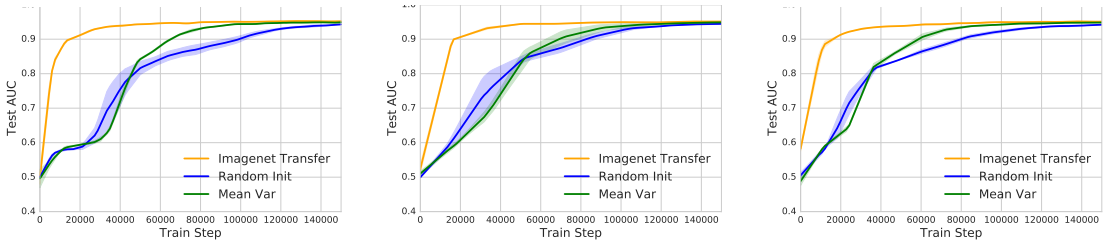
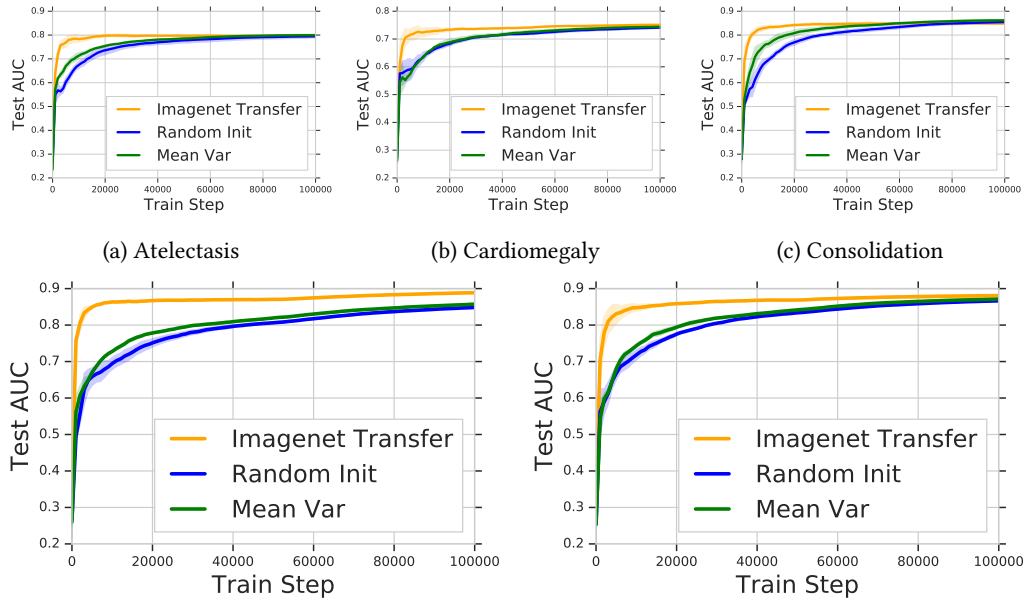


Figure 15: **Comparison of convergence speed for different initialization schemes on Retina with various model architectures.** The three plots present the results for CBR-LargeW, CBR-Small and CBR-Tiny, respectively.

```
kernels = []
for sigma in sigmas:
  for frequency in freqs:
    for theta in range(n_angles):
      theta = theta / n_angles * np.pi
      kernel = np.real(gabor_kernel(frequency, theta=theta,
                                    sigma_x=sigma, sigma_y=sigma))
      kernel_size = kernel.shape[0]
      if kernel_size > kernel_resize:
        kernel = resize(kernel, (kernel_resize, kernel_resize))
        kernel_size = kernel.shape[0]
      else:
        assert kernel_size >= kernel_crop
      # center crop
      size_delta = kernel_size - kernel_crop
      kernel = kernel[size_delta//2:-(size_delta-size_delta//2),
                      size_delta//2:-(size_delta-size_delta//2)]
```



(a) Atelectasis (b) Cardiomegaly (c) Consolidation

(d) Edema (e) Pleural Effusion

Figure 16: **Comparison of convergence speed for different initialization schemes on the CheXpert data with Resnet-50.**

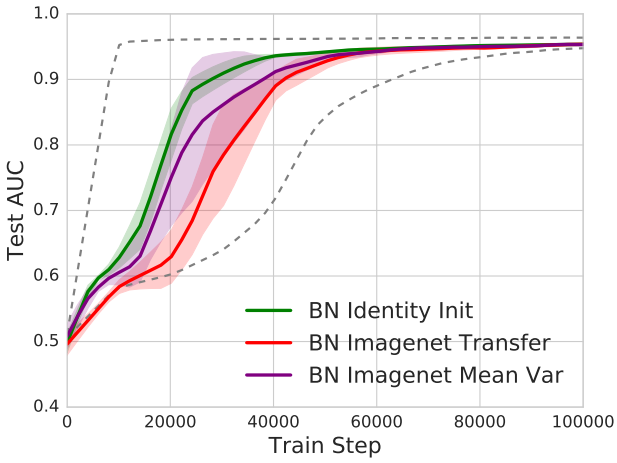


Figure 17: **Comparing different ways of importing the weights and statistics for batch normalization layers.** The rest of the layers are initialized according to the Mean Var scheme. The two dashed lines show the convergence of the ImageNet init and the Random init for references. The lines are averaged over 5 runs.

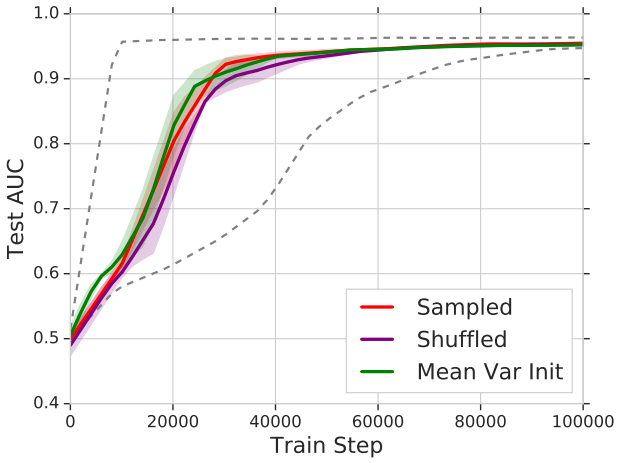


Figure 18: **The Mean Var Init converges with a similar speed to using the full empirical distribution of the pretrained ImageNet weights.** The plots show the convergence speed of initializing by sampling from the empirical ImageNet weight distribution, and initializing by randomly shuffling the pretrained weights (i.e. sampling without replacement). We see that Mean Var converges at a similar speed to using the full empirical distribution. All lines are averaged over 3 runs, and the dashed lines show the convergence of the ImageNet init and the Random init as a reference.

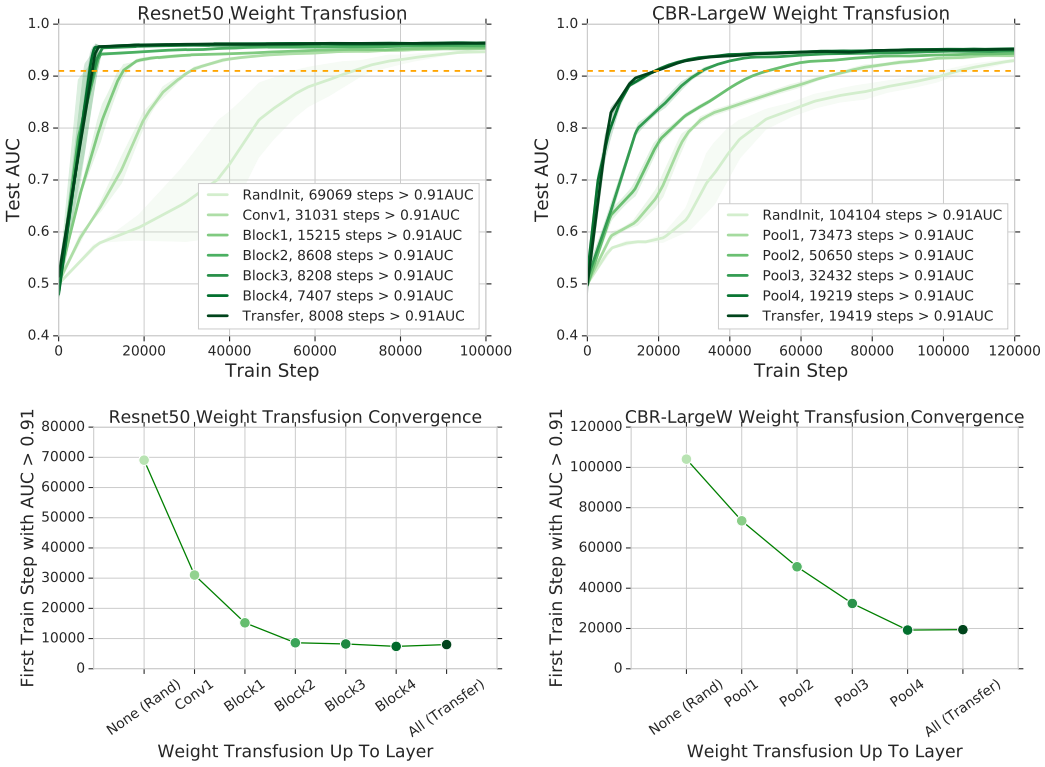


Figure 19: **Weight transfusion results on Resnet50 (from main text) and CBR-LargeW**. These broadly show the same results — reusing pretrained weights for lowest layers give significantly larger speedups. Because CBR-LargeW is a much smaller model, there is slightly more change when reusing pretrained weights in high layers, but we still see the same diminishing returns pattern.

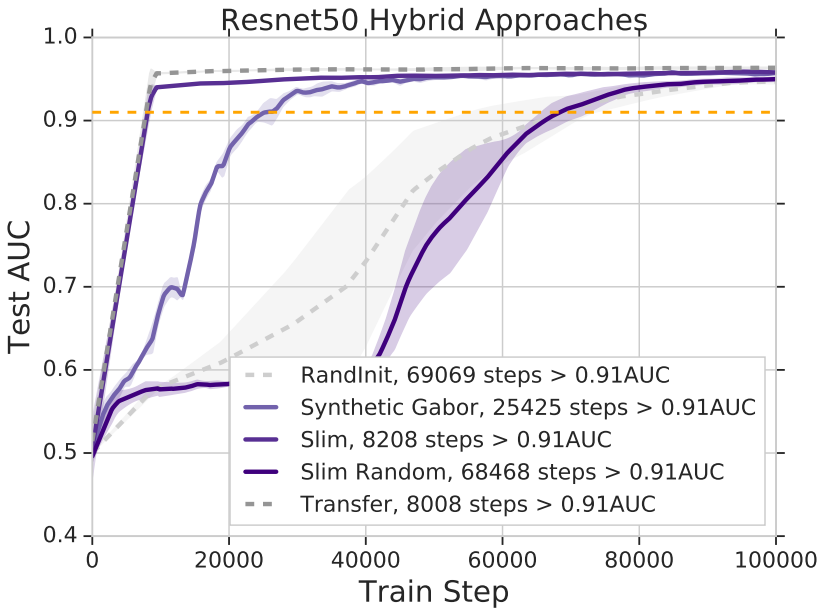


Figure 20: **Convergence of Slim Resnet50 from random initialization**. We include the convergence of the slim Resnet50 — where layers in Block3, Block4 have half the number of channels, and when we don't use any pretrained weights. We see that it is significantly slower than the hybrid approach in the main text.

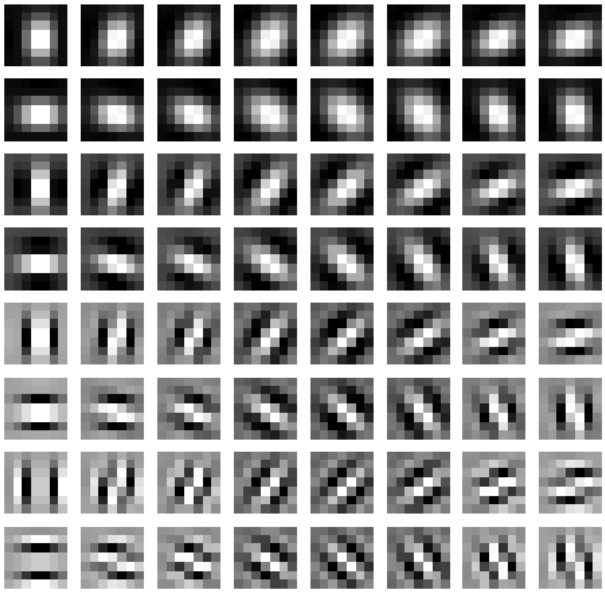


Figure 21: Synthetic Gabor filters used to initialize the first layer if neural networks in some of the experiments in this paper. The Gabor filters are generated as grayscale images and repeated across the RGB channels.

```
        kernels.append(kernel)
    return kernels
```

Figure 21 visualize the synthetic Gabor filters. To ensure fair comparison, the synthesized Gabor filters are scaled (globally across all the filters with a single scale parameter) to match the numerical magnitudes of the learned Gabor filters.