Pre-training on Grayscale ImageNet Improves Medical Image Classification

Yiting Xie* and David Richmond*

IBM, Watson Health, Cambridge MA 02142, USA
daverichmond@gmail.com

Abstract. Deep learning is quickly becoming the de facto standard approach for solving a range of medical image analysis tasks. However, large medical image datasets appropriate for training deep neural network models from scratch are difficult to assemble due to privacy restrictions and expert ground truth requirements, with typical open source datasets ranging from hundreds to thousands of images. A standard approach to counteract limited-size medical datasets is to pre-train models on large datasets in other domains, such as ImageNet for classification of natural images, before fine-tuning on the specific medical task of interest. However, ImageNet contains color images, which introduces artefacts and inefficiencies into models that are intended for single-channel medical images. To address this issue, we pre-trained an Inception-V3 model on ImageNet *after* converting the images to grayscale through a common transformation. Surprisingly, these models do not show a significant degradation in performance on the original ImageNet classification task, suggesting that color is not a critical feature of natural image classification. Furthermore, models pre-trained on grayscale ImageNet outperformed color ImageNet models in terms of both speed and accuracy when refined on disease classification from chest X-ray images.

Keywords: Domain Adaptation · Transfer Learning

1 Introduction

Deep learning algorithms, especially Convolutional Neural Networks (ConvNets), have gained great popularity in the field of medical image analysis in recent years [1]. ConvNet-based algorithms are rapidly replacing traditional machine learning algorithms, based on human-engineered features, for tasks such as image classification [2], object detection [3], and semantic segmentation [4].

There are two general strategies for training ConvNets: (1) training a model from randomly initialized weights, and (2) pre-training a model on a related task, and then refining the model on the target task. The former approach, referred to as "training from scratch", typically requires very large datasets to avoid overfitting and achieve state of the art results. Since, medical datasets are often very small, due to privacy restrictions and the expert knowledge required to generate

* equal contribution

ground truth, transfer learning from a pre-trained model is a popular approach for medical image analysis. There are numerous publicly available models that have been pre-trained on the ImageNet dataset [5], which consists of over 1.2 million labeled photographs. In a recent paper, Rajpurkar et al. [2] fine-tuned a DenseNet model, pre-trained on ImageNet, on a large-scale chest X-ray dataset [6] for a multi-disease classification problem and achieved state-of-the-art results.

However, the choice to start from a pre-trained model has implications for the final ConvNet design. In order to take a ConvNet that was pre-trained on natural images, and fine-tune it on medical images, the medical images need to be pre-processed to conform with the shape and structure of the original color images used to train the network. Since medical images often contain only a single channel, this usually involves stacking each grayscale image to a 3-channel pseudo-color image to mimic the RGB structure of natural images. However, the stacked 3-channel grayscale image does not contain any color information. Therefore, it is unclear whether the filters learned from color images are fully utilized in transfer learning, especially for filters in the first and second layers of the ConvNet, which represent lower level features such as colors and edges.

To address this issue, we trained a ConvNet using a grayscale version of the ImageNet data. A ConvNet was first trained from scratch on grayscale images converted from the ImageNet dataset using a standard transformation [7]. Then the pre-trained ConvNet was fine-tuned on two large-scale chest X-ray datasets for two different tasks: the NIH x-ray dataset [6] for multi-disease classification, and the Indiana University chest x-ray dataset [8] for normal image classification. We demonstrate that a network pre-trained on grayscale ImageNet is a better starting point for transfer learning on medical images, because it (1) leads to more accurate classification performance of the final model, (2) increases the speed of inference, due to the simplified kernel in the first model layer, and (3) removes the need for unnecessary pre-processing before inference.

2 Method

This study consists of two parts: training Inception-V3 models [9] from scratch on ImageNet and then fine-tuning the pre-trained models on the NIH and Indiana X-ray datasets.

Inception-V3 was first trained from scratch on the original color ImageNet dataset (LSVRC2012) to reproduce published state-of-the-art results (see Figure 1(a)). The color ImageNet model was evaluated on the original test set of LSVRC2012. Then the same ConvNet architecture[1] was trained from scratch (using the same optimization parameters) on the same ImageNet dataset (LSVRC2012) after converting the images to grayscale, using the Luma transformation [7] (see Figure 1(b)). The grayscale model was evaluated on the grayscale version of the same test set.

[1] The only difference was that the kernel on the input layer was reduced from 3-channel to 1-channel

The pre-trained color Inception-V3 model was then fine-tuned on both the NIH and Indiana University X-ray datasets for the respective disease classification tasks (see Figure 2(a)). The fine-tuned model was tested on a held-out test set to establish the benchmark performance. Next, the pre-trained grayscale Inception-V3 model was fine-tuned on the NIH and Indiana University X-ray datasets for the same disease classification tasks (see Figure 2(b)). The fine-tuned model was tested on the same held-out test set to compare performance of the two approaches.

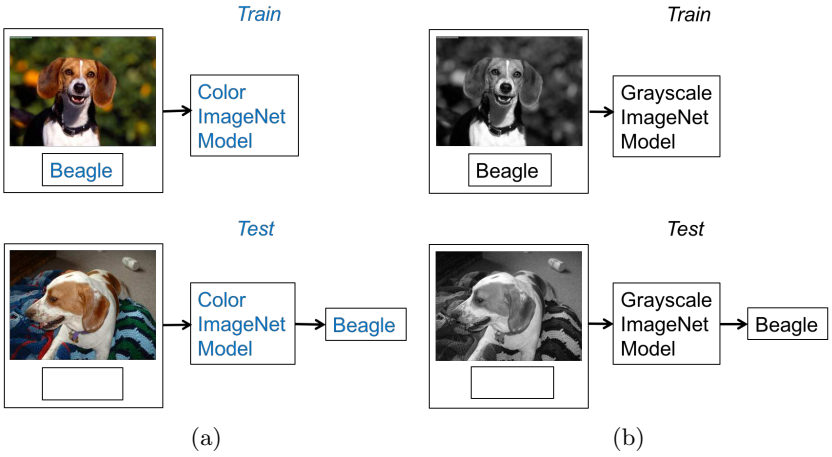


Fig. 1. Training a model from scratch on ImageNet. (a) A 3-channel model is trained and tested on color ImageNet data. (b) A 1-channel model is trained and tested on grayscale ImageNet data.

3 Results

The ConvNet models described in this section were implemented in Tensorflow, and trained using asynchronous Stochastic Gradient Descent (SGD) on two NVIDIA GTX 1080 Ti GPUs.

3.1 Experiment 1: Training color and grayscale models from scratch for ImageNet-based classification

Training and validation images were from the ImageNet Large Scale Visual Recognition Challenge 2012 (LSVRC2012 [5,10]). In total, 1,281,167 images were used for training and 50,000 images for validation. For the classification challenge, there are 1000 image categories and each image belongs to one category.

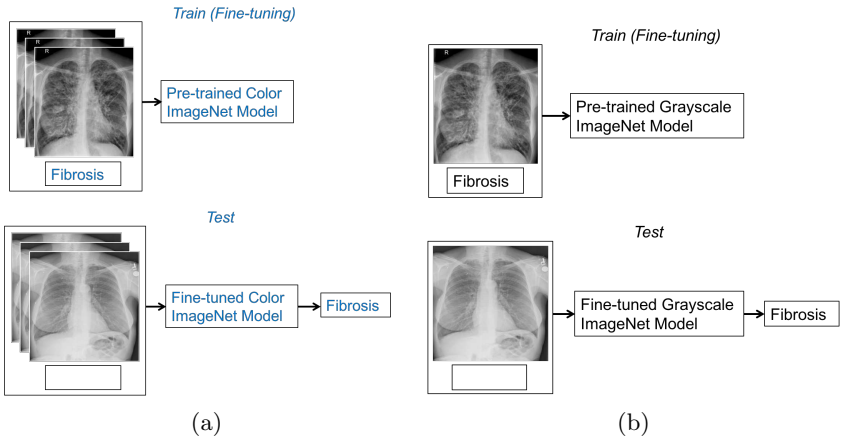


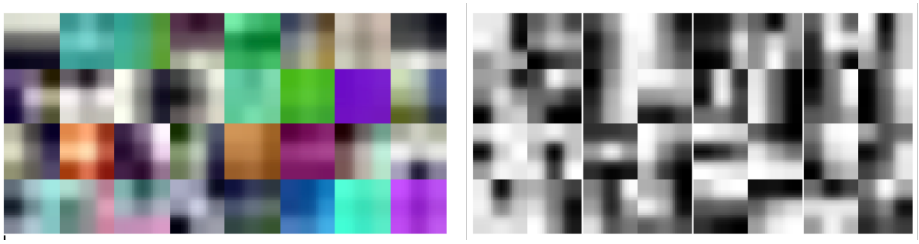
Fig. 2. Fine-tuning the ImageNet-trained models on X-ray data. (a) The color model is fine-tuned and tested on X-ray data after converting the X-ray images to 3-channel pseudo-color images. (b) The grayscale model is fine-tuned and tested on X-ray data without any image transformation.

For training on the color ImageNet data, standard augmentation methods [11] were used: cropping images based on a distorted version of the annotated bounding box, random horizontal flipping, and altering the intensities of the RGB channel. RMSProp optimizer was used with a decay factor of 0.9. The initial learning rate was set to 0.01 with a decay factor of 0.94 every 2 epochs [9]. The batch size was set to 64. The network was trained until the loss converged. The model converged after about 14 days and 1.67 million steps (around 84 epochs). The validation accuracy was 0.9169 for top-5 and 0.7372 for top-1. The state-of-the-art validation accuracy using Inception-V3 on the same dataset is 0.939 for top-5 and 0.780 for top-1 [12].

For training on the grayscale ImageNet data, the same hyper-parameters and augmentation methods were used. After augmentation, the color images were converted to grayscale, using the Luma transformation [7]. The batch size was set to 64 and the network was trained until the loss converged. The model converged after about 16 days and 1.92 million steps (around 100 epochs). The validation images were also converted to grayscale, and the validation accuracy was 0.9117 for top-5 and 0.7323 for top-1. Surprisingly, the performance of the model trained and tested on grayscale ImageNet was only 0.5% lower than the color model, suggesting that color is not a critical feature in image classification. The results are summarized in Table 1. Figure 3 shows the first-layer kernels learned from the color model and the grayscale model.

Table 1. Evaluation results on ImageNet classification

	Top-5 Accuracy	Top-1 Accuracy
Color	0.9169	0.7372
Grayscale	0.9117	0.7323



(a) (b)

Fig. 3. First-layer kernels learned by training on (a) color ImageNet, and (b) grayscale ImageNet.

3.2 Experiment 2: Fine-tuning on NIH X-ray dataset

The NIH X-ray dataset consists of 112,120 frontal chest X-ray images from more than 30,000 patients. There is a total of 14 lung diseases in this dataset: Atelectasis, Cardiomegaly, Emphysema, Effusion, Hernia, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Fibrosis, and Pleural thickening. Each X-ray image could contain any number of the 14 diseases, or no finding. In total, 60,361 images have no findings. The image disease labels were mined from radiological reports using natural language processing and released together with the X-ray images.

The X-ray dataset was partitioned into 3 subsets for training, validation and testing following the same strategy used by Wang et al. [6]: 70% for training, 10% for validation, and 20% for testing. Since the same patient could have multiple X-ray images in this dataset, partition was performed to ensure that there were no overlapping patients among the 3 subsets. Each X-ray image was down-sampled to a fixed size compatible with the network input, and in the case of the color model, stacked to form a 3-channel image. Random horizontal flipping was used for training data augmentation. The final fully-connected layer in the pre-trained model was replaced with a fully connected layer producing a 14-label output. A sigmoid nonlinearity was used and the final output was the disease probabilities for the 14 disease classes. The learning rate was set to 0.0001 and the batch size was set to 32. The network was trained end-to-end until the validation loss converged. For evaluation, the Area Under the ROC Curve (AUC) was calculated for the 14 diseases on the testing subset.

The color model converged after about 85k steps. The grayscale model converged after about 250k steps. The AUC values on the test dataset is summarized in Table 2. The p-values were also computed to assess the statistical significance

of the differences between each pair of ROC curves [13]. When comparing the grayscale vs color model, all 14 categories had improved performance, and 8 out of 14 categories had performance improvements that were statistically significant. Furthermore, the grayscale model was approximately 20% faster for inference than the color model (see Table 4).

Table 2. Results (AUC) on NIH X-ray test data for 14 diseases after fine-tuning a pre-trained Inception-V3 model (c=color, g=grayscale). * p<0.05, ** p<0.01

Disease	**Avg**	**Atelectasis**	**Cardiomegaly**	**Emphysema**	**Effusion**	**Hernia**	**Infiltration**
AUC (c)	0.7498	0.7613	0.7785	0.7898	0.8320	0.6843	0.6835
AUC (g)	0.7706	0.7824**	0.8091*	0.8393**	0.8423*	0.7035	0.6895
Mass	**Nodule**	**Pneumonia**	**Pneumothorax**	**Consolidation**	**Edema**	**Fibrosis**	**PT**
0.7132	0.6807	0.6905	0.8145	0.7489	0.8689	0.7310	0.7204
0.7498**	0.7096**	0.7021	0.8326*	0.7606	0.8784	0.7454	0.7452*

3.3 Experiment 3: Fine-tuning on Indiana University X-ray dataset

The Indiana University X-ray dataset consists of around 8000 X-ray images from more than 3000 different patients. There are more than 100 types of disease labels in this dataset as well as the label indicating whether the image is normal or not. In our experiment, one frontal X-ray image was selected for each patient with associated medical report, resulting in a total of 3691 patients and images. A binary classification task of normal versus abnormal was performed on this dataset.

The Indiana University X-ray dataset was partitioned into 70%, 10%, and 20% for training, validation, and testing. Random horizontal flipping was used for training augmentation. The softmax loss was used and the final layer output was the probability indicating whether the image was normal or not. The learning rate was set to 0.0001 and the batch size was set to 32. The network was trained end-to-end until the validation loss converged. For evaluation, the accuracy and the Area Under the ROC Curve (AUC) was calculated on the testing subset.

The color model converged after about 11k steps. The grayscale model converged after about 9k steps. The accuracy and AUC values on the test dataset is summarized in Table 3. For the Indiana X-ray dataset, the grayscale model had improved performance; however, the difference in performance was not statistically significant. Inference with the grayscale model was approximately 5% faster (see Table 4).

4 Discussion and Conclusion

Due to the limited size of most medical imaging datasets, pre-training ConvNet models on large image repositories, such as ImageNet, is a common initialization

Table 3. Results on the Indiana X-ray test data for normal vs abnormal classification using pre-trained Inception-V3 model.

	Accuracy	AUC
Color	0.7225	0.7124
Grayscale	0.7262	0.7285

Table 4. Inference time (ms/image)

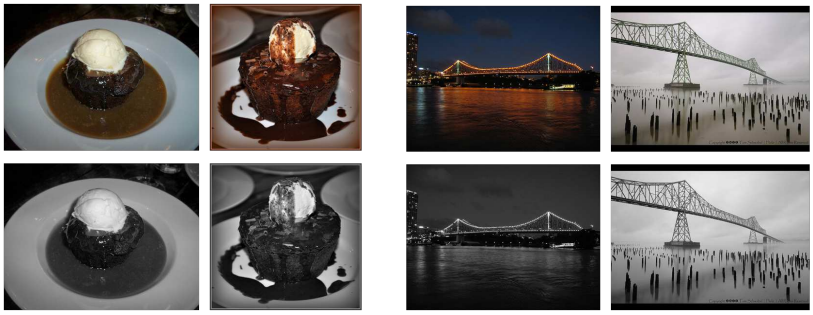
	NIH	Indiana University
Color	8.0	7.5
Grayscale	6.4	7.1

strategy. However, due to the long training time required to train models from scratch on ImageNet, they are typically downloaded from publications focused on processing natural images. This leads to artefacts, whereby single-channel medical images must be pre-processed to 3-channel pseudo-color images before they can be analyzed by the network.

We show that transferring ImageNet data to a single-channel (i.e., grayscale) domain leads to better pre-trained models that (1) achieve higher classification accuracy after being fine-tuned on medical X-ray image data, (2) are faster during inference, and (3) avoid unnecessary pre-processing. We hypothesize that the network pre-trained on grayscale images has the potential to learn more features relevant to grayscale images, which serves to boost the transfer learning performance when applied to a grayscale medical dataset.

Surprisingly, after converting both training and testing sets of the ImageNet LSVRC2012 data to grayscale, the test set performance was only reduced by 0.5%, from a top-5 accuracy of 0.9169 for the color model to 0.9117 for the grayscale model. This result was counter to our expectation that color would be an important feature for accurate classification of natural images. However, it seems to be consistent with the success of colorization methods [14,15] which produce realistic-looking color images from grayscale image information.

We also compared class-specific performance between the two models. While for the majority of the classes, the two models had very similar performance, the color model outperformed grayscale model on classes such as ice-cream and mink. For example, the grayscale model classified some ice-cream images into the chocolate sauce class. The grayscale model performed better on classes including pier and printer (the color model classified a lot of pier images into suspension bridges). An intuitive explanation could be that color information is more important for discriminating between certain classes such as ice cream vs chocolate sauce (e.g., chocolate sauce is brown) but not for other classes such as pier vs suspension bridge. Figure 4 shows some example images from these two classes in color and in grayscale.



(a) (b)

Fig. 4. Example images belonging to (a) the ice-cream class and (b) the pier class. Upper row shows the color images and lower row shows the corresponding grayscale images.

In conclusion, color does not seem to be a critical feature for accurate classification of natural images, and pre-training on grayscale images can give a boost in both speed and accuracy when fine-tuning on medical images. In future, it would be interesting to apply this approach to semantic segmentation and object detection in medical images, through the use of standard network architectures such as fully convolutional networks (FCN), and region-based convolutional neural networks (R-CNN). It would also be interesting to explore additional image transformations that may be more appropriate for different imaging modalities, such as Ultrasound and Magnetic Resonance Imaging.

References

1. G. Lijens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sanchez: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
2. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, `https://arxiv.org/pdf/1711.05225.pdf`.
3. B. van Ginneken, A. A. Setio, C. Jacobs, and F. Ciompi: Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: IEEE ISBI, pp. 286–289 (2015)
4. J. Wang, J. D. MacKenzie, R. Ramachandran, and D. Z. Chen: A deep learning approach for semantic segmentation in histology tissue images. In: MICCAI, pp. 176–184 (2016)
5. ImageNet, `http://www.image-net.org/`.
6. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers: ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: IEEE CVPR, pp. 2097–2106 (2017)

7. Luma, https://en.wikipedia.org/wiki/Luma_%28video%29.
8. Open-i Biomedical Image Search Engine, https://openi.nlm.nih.gov/faq.php.
9. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. and Wojna: Rethinking the Inception Architecture for Computer Vision. In: CVPR, pp. 2818–2826 (2016)
10. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. Li: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015)
11. A. Krizhevsky, I. Sutskever, and G. E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks. In: NIPS, pp. 1097–1105 (2012)
12. TensorFlow-Slim image classification model library, https://github.com/tensorflow/models/tree/master/research/slim.
13. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**(3), 837–845 (1988)
14. G. Larsson, M. Maire, and G. Shakhnarovich: Learning Representations for Automatic Colorization. In: ECCV, pp. 577–593 (2016)
15. R. Zhang, P. Isola, and A. A. Efros: Colorful Image Colorization, https://arxiv.org/pdf/1603.08511.pdf.