Self-Supervised learning for cancer image classification with limited data

H. Haggerty[1] Dr. R. Chandra[2]

[1]School of Mathematics and Statistics
UNSW

[2]School of Mathematics and Statistics
UNSW

June 25, 2023

Table of Contents

1. Motivation
2. Review of transfer learning
3. What is Self-Supervised learning?
4. Framework
5. Results
6. Discussion

Table of Contents

1. Motivation
2. Review of transfer learning
3. What is Self-Supervised learning?
4. Framework
5. Results
6. Discussion

Motivation

- Cancer is the second leading cause of death worldwide.
- Early diagnosis is often *the* determining factor in prognosis.
- Deep learning achieves excellent results in cancer image classification (e.g. for skin, lung etc) *given* a sufficient amount of labelled data.
- **Problem**: For many kinds of cancer, large labelled datasets do not exist. This is true for several types of cancer with increasing prevalence in the developing world (e.g. oral cancer) which is of particular concern. Moreover, for rare forms of cancer, large labelled datasets will likely *never* exist. Developing more data efficient diagnostic techniques is therefore essential.

A different approach

- Cancer image classification usually involves transfer learning: initialise the network weights with that of a network pretrained on ImageNet in a **supervised fashion**.
- This raises the question: can we do better? *Yes!* Through a different pretraining mechanism known as **self-supervised learning**. We demonstrate this for skin lesion images.

Table of Contents

1. Motivation
2. Review of transfer learning
3. What is Self-Supervised learning?
4. Framework
5. Results
6. Discussion

What is transfer learning?

- Instead of randomly initialising neural network weights **reuse weights** from a network trained on a related task.
- Usually the weights come from a network trained on supervised ImageNet classification.
- Has been demonstrated that transfer learning in this way can lead to significant performance gains compared to training from random initialisation. The advantage is particularly large when the amount of labelled data available is low.

How to perform transfer learning

Assume we have a labelled dataset $\mathcal{D}$ and a model $L \circ f_\theta$. f_θ is a pretrained backbone model, L is a randomly initialised linear layer. We are going to perform supervised learning.

- Naively, you could just train the network the usual way, i.e. as though both L and f_θ were randomly initialised. Many papers still do this even though it is suboptimal.
- **Problem**: L is randomly initialised and f_θ is not. This is a problem because the pretrained features get harmed at the start of training by aligning f_θ with L (with respect to $\mathcal{D}$).
- **Solution**: freeze the encoder f_θ for several epochs, while only updating the head L. This is essentially multinomial logistic regression on $x' = f_\theta(x)$ with θ frozen. Then unfreeze the backbone encoder and continue training as usual.

Table of Contents

1. Motivation
2. Review of transfer learning
3. What is Self-Supervised learning?
4. Framework
5. Results
6. Discussion

What is Self-Supervised learning?

Self-supervised learning is a way of pretraining models, but *without* using label information. This has had huge success in NLP, but less so in computer vision until recently.

What is Self-Supervised learning?

The goal is the same as in standard (supervised) pretraining: to learn a model initialisation of f_θ. The difference is no label information is used in the pretraining phase.

Supervised learning vs. Self-Supervised learning

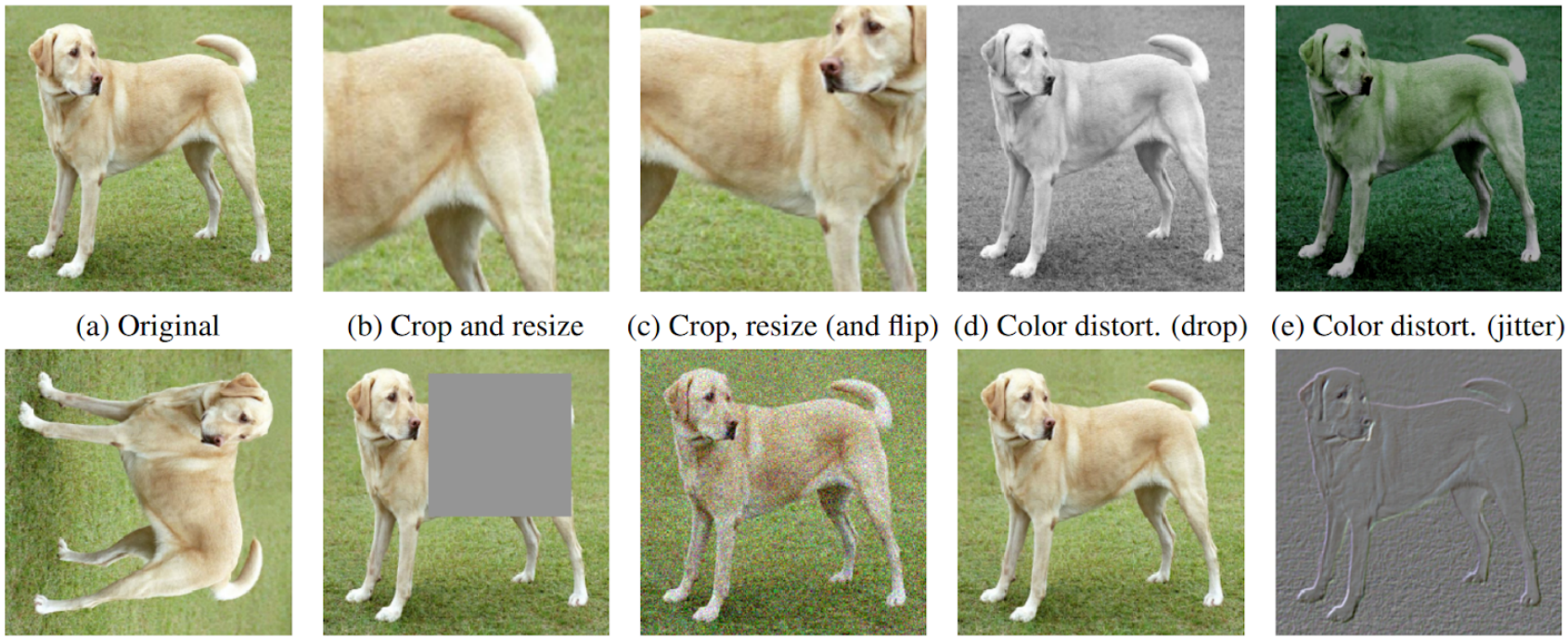
Give the model several examples of each category.

Cat

Dog

Self-Supervised learning

We don't have labels. Instead, tell the model these images are the *same* (in some sense). In other words, for two distorted views of the same image, $t(x), t'(x)$, we want $f_\theta(t(x)) \approx f_\theta(t'(x))$. This is called *invariance*.

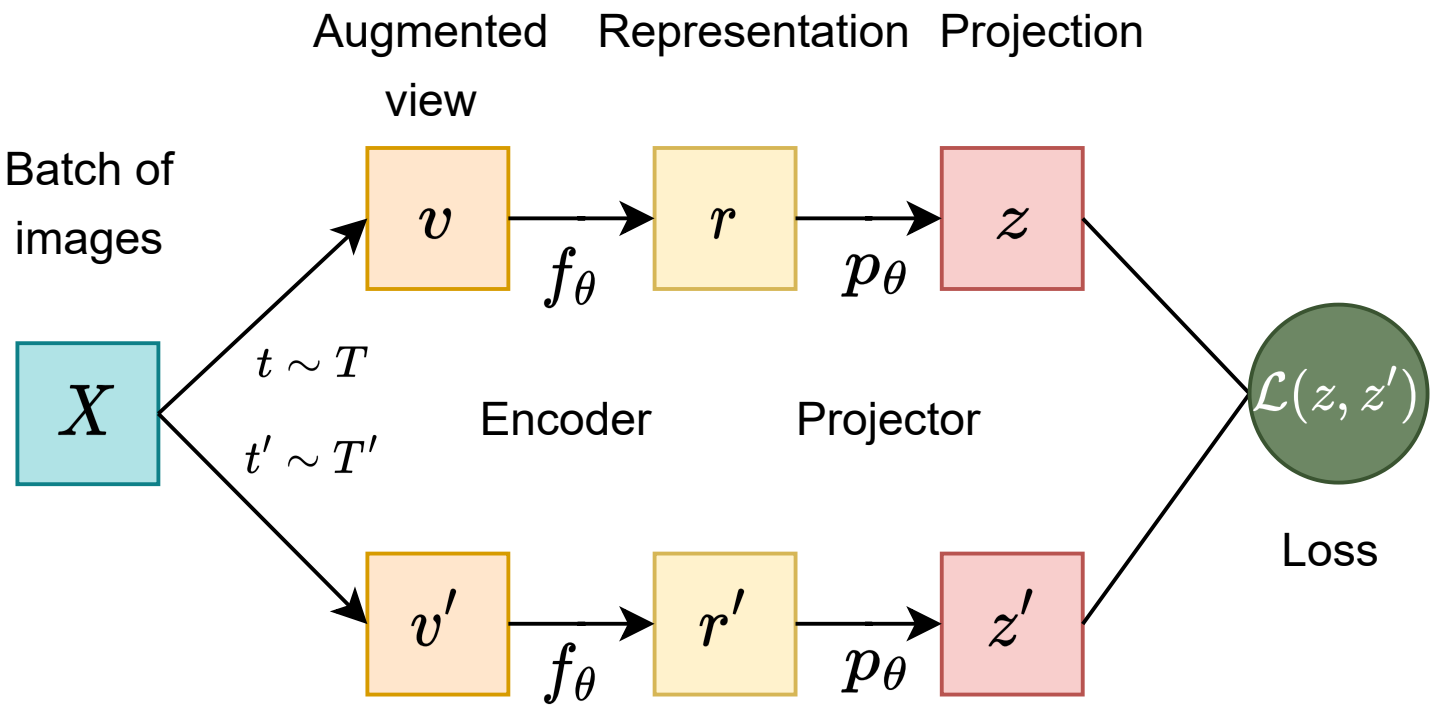


(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Joint embedding architecture

This is done using a *joint embedding architecture.* Key point: use data augmentation to produce two distorted views of a batch. Pass through a neural network to get a d dimensional representation of each image. Loss function enforces similarity.



Joint embedding architecture

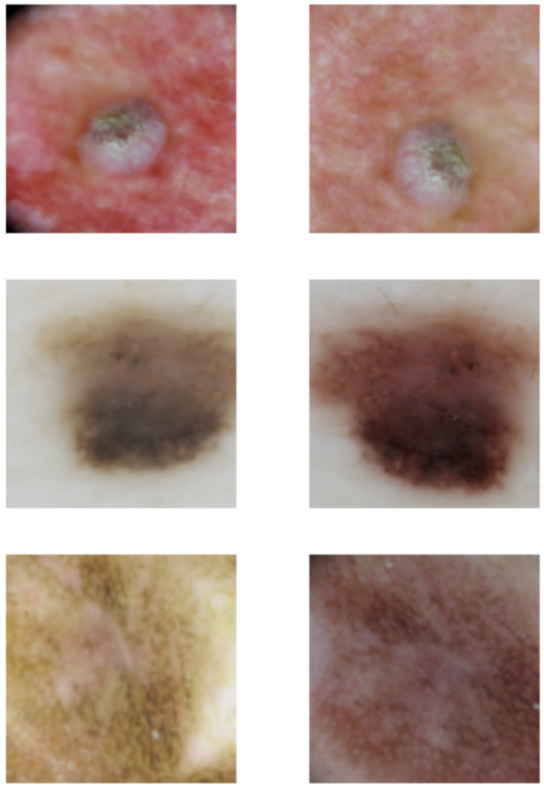


Figure: Batch of size 3. Each row has two distorted views of the same sample. Pass both views through neural network to get representation. Loss function makes representations across rows similar.

Barlow Twins

- Simple way of writing down loss function
- Compute cross correlation between branches
- Loss tries to equate cross correlation to identity matrix
- Key point: loss has two terms, invariance term and redundancy reduction term, which you can think of as a regularisation term:

$$\mathcal{C} = \frac{1}{n}(z')^\top z,$$
$$\mathcal{L}_{\mathcal{BT}} = \sum_i (\mathcal{C}_{ii} - 1)^2 + \lambda \sum_{i \neq j} \mathcal{C}_{ij}^2,$$

First term is invariance term, second term is regularisation term. λ is a hyperparameter.

Table of Contents

1. Motivation
2. Review of transfer learning
3. What is Self-Supervised learning?
4. Framework
5. Results
6. Discussion

Data for our work

Lesion type	Train	Test
Actinic Keratosis	306	498
Basal Cell Carcinoma	500	2549
Benign Keratosis	467	1663
Dermatofibroma	55	173
Melanoma	500	3339
Nevus	500	10601
Squamous Cell Carcinoma	171	414
Vascular Lesion	55	186
Total	2554	19423

Table: Number of samples in the training and test set, per lesion type.

- Labelled skin lesion dataset.
- 8 lesion categories, 4 benign and 4 malign.
- Only 171 squamous cell carcinoma training samples - very challenging dataset.

Framework

- We are going to compare supervised pretraining with self-supervised pretraining.
- Same architecture for both: ResNet-50, denoted f_θ.
- f_θ either pretrained in supervised manner on ImageNet or self-supervised (with Barlow Twins) also on ImageNet.
- Therefore model initialisation is: $L \circ f_\theta$.

Algorithm 1 Transfer Learning

1: Freeze the encoder f_θ.
2: Train the final linear layer L for one epoch.
3: Unfreeze f_θ.
4: Run a learning rate finder to find a good maximal learning rate lr.
5: Train network for specified number of epochs using 1cycle policy and lr: `fit_one_cycle(epochs, lr)`.

1cycle policy

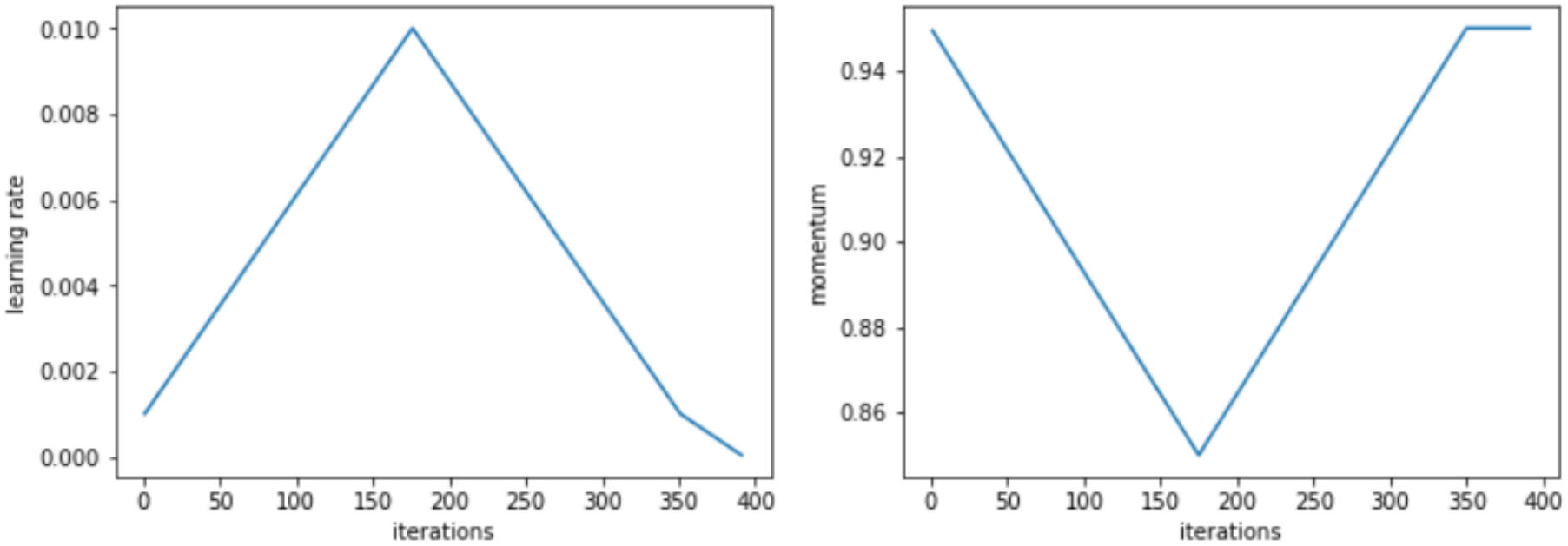


Figure: 1cycle policy. Learning rate and momentum follow inverse schedules.

Table of Contents

1. Motivation
2. Review of transfer learning
3. What is Self-Supervised learning?
4. Framework
5. Results
6. Discussion

Results

Self-supervised pretraining was superior!

	Supervised	**Self-supervised**
Mean Accuracy	0.663	0.703
Standard Deviation	0.0198	0.014
Number of Models	35	35

Table: Model fine tuning results for two kinds of initial weights.

- Also superior on other metrics, for example F1 score.
- Self-supervised models had higher (mean) F1 score for every lesion category.
- Largest percentage difference was for squamous cell carcinoma, which also had the lowest training support, with self-supervised F1 score 10.8% higher than supervised.

Results

Lesion Type	f1-score (sup)	f1-score (self-sup)	Support
Actinic keratosis	0.44	0.45	498
Basal cell carcinoma	0.72	0.73	2549
Benign keratosis	0.48	0.49	1663
Dermatofibroma	0.33	0.37	173
Melanoma	0.56	0.57	3339
Melanocytic nevus	0.77	0.82	10601
Squamous cell carcinoma	0.35	0.39	414
Vascular lesion	0.31	0.51	186
Weighted avg.	0.68	0.71	19423
Accuracy	0.66	0.70	19423

Table: Combined classification report for skin lesion classification model. Each value is a mean across 35 runs.

Results

Results are still not acceptable for clinical applications. Two ways to improve self-supervised models:

1. Bigger model
2. More unlabelled data

Bigger model

We tried fine tuning a ResNet-200 (x2) which has over 10x more parameters than ResNet-50. This model had similarly been SSL pretrained on ImageNet. Accuracy was unchanged.

Conclusion: Self-supervised pretraining of bigger models on ImageNet alone does not appear to help medical transfer classification.

More unlabelled data

Next we considered pretraining with Barlow Twins on the target data. The train and test sets are combined for 21977 **unlabelled** samples, and we pretrained a network using Barlow Twins. The initial weights were those from the network that had already been pretrained with Barlow Twins, on ImageNet.

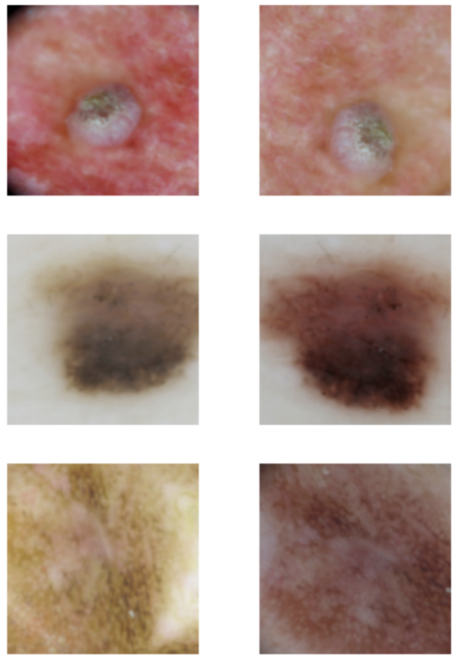


Figure: Barlow Twins augmentations. Each row has two distorted views of the same sample.

Results

	Supervised	**Self-supervised**	**Pre-pretrained**
Mean Accuracy	0.663	0.703	0.723
Standard Deviation	0.0198	0.014	0.0037
Number of Models	35	35	6

- Pre-pretrained models have been pretrained on all unlabelled data, starting from the self-supervised (Barlow Twins) initial weights.
- Weights are pretrained twice: once on ImageNet, then on the current unlabelled data.

Discussion

Our results are for skin lesion images, but we expect results to generalise to other forms of cancer, e.g. oral cancer. Note that over 90% of oral cancers are squamous cell carcinomas. Therefore for similar problems we recommend:

- considering self-supervised pretraining instead of supervised pretraining;
- collecting more unlabelled target data and pretraining a second time before fine tuning.

Table of Contents

1. Motivation
2. Review of transfer learning
3. What is Self-Supervised learning?
4. Framework
5. Results
6. Discussion

Discussion

- Recall that both base models had been pretrained on the same data, i.e ImageNet.
- ImageNet features 1.2 million samples: tiny relative to web image data.
- Obtaining more data for self-supervised pretraining therefore *much* easier than for supervised pretraining (annotating ImageNet alone was a challenge).
- We found self-supervised pretraining is *already* superior, even when the pretraining data is the same i.e. ImageNet.
- Pretraining self-supervised methods on larger datasets may yield significant gains.

References I

[1] Jacques Ferlay et al. "Cancer statistics for the year 2020: An overview". In: *Int J Cancer* (2021). Online ahead of print.

[2] International Skin Imaging Collaboration. *ISIC Archive: A Comprehensive Resource for Skin Imaging Data*. Accessed: 2023-04-20. 2023. URL: https://www.isic-archive.com.

[3] Yann LeCun and Ishan Misra. *Self-supervised learning: The dark matter of intelligence*. Facebook AI Blog. 2021.

[4] Jure Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: *International Conference on Learning Representations (ICLR)*. 2021.

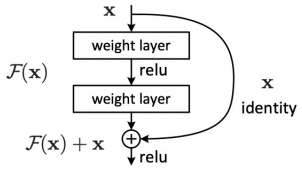
[5] Fuzhen Zhuang et al. "A Comprehensive Survey on Transfer Learning". In: *Proceedings of the IEEE* 109.1 (2021), pp. 43–76.

Thank you for listening to my talk.

PRETRAINING → PREPARATION → FINE TUNE → RESULTS → CLINICAL IMPLICATIONS

ImageNet
(natural images)


ResNet CNN architecture



Supervised learning
Train ResNet CNN on ImageNet using label information (supervised).

Returns:
(encoder,linear)

Barlow Twins
Train ResNet CNN on ImageNet *without* using label information (self-supervised).

Returns:
(encoder,projector)

Small labelled skin cancer dataset

Cut ResNet at penultimate layer (leaves encoder)

Append head to encoders

Freeze encoders

Train head for 1 epoch

Unfreeze encoders

Run learning rate finder

Train models with 1cycle policy

Labelled test set

Compare models: F1, accuracy

Open source software

Future work

Compare methodology on other small medical image datasets.

Improved diagnostic capacity in low data scenarios

Particular relevance to developing world

Most studies use supervised pretraining. Potential for all such work to be enhanced via self-supervised pretraining and learning rate schedule.