Deep Cascade Learning for Optimal Medical Image Feature Representation

Junwen Wang* JW7U18@SOTON.AC.UK
University of Southampton
Hampshire, Southampton, UK

Xin Du* XD3Y15@SOTON.AC.UK
University of Southampton
Hampshire, Southampton, UK

Katayoun Farrahi K.FARRAHI@SOTON.AC.UK
University of Southampton
Hampshire, Southampton, UK

Mahesan Niranjan MN@ECS.SOTON.AC.UK
University of Southampton
Hampshire, Southampton, UK

Abstract

Cascade Learning (CL) is a new and alternative form of training a deep neural network in a layer-wise fashion. This varied training strategy results in different feature representations, advantageous due to the incremental complexity induced across layers of the network. We hypothesize that CL is inducing coarse-to-fine feature representations across layers of the network, differing from traditional end-to-end learning, advantageous for medical imaging applications. We use five different medical image classification tasks and a feature localisation task to show that CL is a superior learning strategy. We show that transferring cascade learned features from cascade trained models from a subset of ImageNet systematically outperforms transfer from traditional end-to-end training, often with statistical significance, but never worse. We demonstrate visually (using Grad-CAM saliency maps), numerically (using granulometry measures), and with error analysis that the features and also errors across the learning paradigms are different, motivating a combined approach, which we validate further improves performance. We find the features learned using CL are more closely aligned with medical expert labelled regions of interest on a large chest X-ray dataset. We further demonstrate other advantages of CL, such as robustness to noise and improved model calibration, which we suggest future work seriously consider as metrics to optimise, in addition to performance, prior to deployment in clinical settings.

1. Introduction

While most deep neural network (DNN) training start with an arbitrarily fixed architecture, researchers have also explored adapting network architectures to the complexity of a problem. Adaptive training has the advantage of training deeper networks with limited resources, since the features can be cached at any point in time. Examples include

* Equal Contribution

the pioneering work of resource allocating networks (Platt, 1991), and their function approximation and probabilistic variants (Kadirkamanathan and Niranjan, 1993; Roberts and Tarassenko, 1994). A powerful member of this family of approaches is the cascade correlation algorithm (Fahlman and Lebiere, 1990). Deep Cascade Learning (**CL**) (Marquez et al., 2018), builds on cascade correlation algorithm, as an alternative way of training a DNN. The motivations are to match architecture complexity to problem complexity in a constructive way as well as extracting coarse-to-fine representations. This learning paradigms differs from traditional end-to-end (**E2E**) learning, whereby all of the layers of the network are learned simultaneously, resulting in varied feature representations. Recent studies have demonstrated, using information bottleneck theory, that CL has advantages with respect to learning when to stop training a network (Du et al., 2021). Further work (Du et al., 2019; Stumpf et al., 2020) demonstrates the superior performance of CL on downstream tasks. Other learning paradigms, such as the Gradient Isolated Learning methods (Löwe et al., 2019; Nøkland and Eidnes, 2019; Wang et al., 2021) are also being developed, though our interest is on feature representation for medical imaging that can scale as opposed to local learning and unsupervised methods.

DNNs remain state of the art for computer vision, including medical imaging, since Alexnet championed the ImageNet LSVRC-2010 classification challenge (Krizhevsky et al., 2012). While DNNs excel in terms of performance, they do suffer from being difficult to interpret due to their complexity in addition to being poorly calibrated (Guo et al., 2017), and unstable to noise corruption (Hendrycks and Dietterich, 2019). Interpretability, the ability of a human to understand the link between the features learned and the predictions made, is crucial to gain trust in these systems as they become integrated into clinical settings. Image saliency maps (Simonyan et al., 2014) are one method of interpreting DNNs by highlighting areas of an image to which the output decisions are most sensitive. In radiology, saliency maps can be integrated into the workflow, allowing easy fusion with patient images and computer-generated results (Reyes et al., 2020). In recent work, Grad-CAM (Selvaraju et al., 2020) generates gradient-based saliency maps that enable visualisation of every DNN layer. Assessing the uncertainty of a model's results can also be used to enhance interpretability by understanding which images or areas of an image the model identifies as being difficult (Reyes et al., 2020). Uncertainty estimates can also be viewed as system verification methods; they can be used as a proxy for trust in a system, as a radiologist can verify the confidence levels of a model's predictions. In addition to interpretability, and the importance of reliable uncertainty estimates, noise robustness is also important in the medical context. In real clinical settings, the existence of noise is inherent in the data capture process (Gravel et al., 2004). Sources of noise can be from hardware, algorithm design and parameter settings (Zhang et al., 2020).

In the medical domain, the precondition of large labelled datasets is not always feasible as medical image data often comes from small disease populations, requires costly expert labelling, and has potential privacy implications, motivating research on transfer learning (Bar et al., 2015; Chen et al., 2015; Schlegl et al., 2014; Tajbakhsh et al., 2017; Van Ginneken et al., 2015; Wang et al., 2017). There are two directions that deep learning based transfer learning has been explored in the medical domain. The first set of works (Arevalo et al., 2015; Bar et al., 2015; Carneiro et al., 2015; Chen et al., 2015; Shin et al., 2015; Van Ginneken et al., 2015) transfers knowledge using pre-trained networks from natural images as a feature

extractor, while the majority of the convolutional layers are fixed during transfer on the target domain. The second set of works (Margeta et al., 2017; Shin et al., 2016; Tajbakhsh et al., 2017) fine-tune on the downstream task. In particular, Tajbakhsh *et al.* (Tajbakhsh et al., 2017) fine-tune in a layer-wise fashion on the medical domain. Raghu *et al.* (Raghu et al., 2019) report a critical appraisal of transfer learning from E2E networks, showing little or no improvement on two benchmark medical image tasks. An opposite conclusion is obtained by Ke *et al.* (Ke et al., 2021), demonstrating that the family of architectures, as opposed to the model size, determines performance, showing performance improvements using ImageNet pre-training for transfer. Recently, transferring features via *self-supervised learning* in the source domain has been considered (Azizi et al., 2021; Ericsson et al., 2021; Goyal et al., 2019; Truong et al., 2021). All of the methods mentioned above are transferring features from networks that have been trained in an end-to-end fashion.

In this paper, we demonstrate for the first time the superior feature representations learned using CL, considering five medical imaging classification tasks and one localisation task. We consider the downstream task of transfer learning, where CL shows particularly promising results, outperforming traditional methods. Particularly statistically significant improvements in low data regimes are observed. We even show comparable performance on large data regimes, motivating transfer as a resource efficient strategy rather than training a large network from scratch. Our motivation comes from the nature in which CL works, resulting in a varied feature representation strategy to E2E training. Whereas in E2E training, representations learned in layers across the network depend largely on the trajectory along which the error function is minimized, in CL there is a progression in complexity as more and more layers are added. Thus, one could expect that early layers, trained as low capacity networks, are able to absorb coarse features while later layers will be trained to extract specific features of the problem. This motivates potentially better transferability from early layers of the network, which we confirm in our experiments.

We further quantify the representational difference between CL and E2E training in terms of: *feature localisation* and *robustness*. We show both visually and numerically that the features learned by TCL are more localised to the discriminative target area of interest. We show the errors made by TCL differ to TE2E, motivating a combined approach which outperforms the individual approaches on every dataset, resulting in state of the art medical imaging classification performance. We demonstrate that TCL saliency maps have high granulometry, indicating more concentrated feature activation (E2E saliency maps tend to be less focused). Finally, we show the superiority of TCL in terms of robustness to noise as well as offering better model calibration, and therefore, better uncertainty estimates. In summary, we are showing strong evidence for the use of CL for better medical feature representations.

1.1. Generalizable Insights

The generalizable insights our particular approach and empirical work reveal are as follows:

- We demonstrate that a CL strategy (as opposed to traditional E2E learning), has improved feature representations in DNNs for medical imaging tasks and we recommend that CL be considered by researchers and clinicians in this domain.

- We consider the problem of **how to learn optimal features** for medical imaging and find that a combined strategy of TCL and TE2E will result in optimal performance on five medical imaging downstream tasks (see Figure 2).
- We consider the problem of **when to transfer** for medical imaging. Our results in Figure 2 demonstrate that for small data regimes, TCL will perform as good as or better than TE2E or learning from scratch.
- We demonstrate that CL learns different features, with coarser features in early layers and finer features in later layers (see Figure 5) whereas end-to-end learned features have more evenly distributed granulometry across layers. The features learned with CL are therefore different and we show they are better suited for other factors, such as localisation and robustness.
- We demonstrate that CL provides superior representations for several **types of medical images** (chest X-rays radiography, histopathology, dermatoscopy and endoscopy).
- We demonstrate the superiority of our CL framework in terms of feature localisation using feature visualisation techniques, granulometry results, as well as with experiments on a expert labelled dataset.
- We suggest to evaluate systems beyond simple classification performance and consider issues such as interpretability, calibration as well as robustness when presenting machine learning models to the medical community.

2. Methodology

2.1. Deep Cascade Learning

Cascade Learning (**CL**) (Marquez et al., 2018), illustrated in Figure 1, is an iterative approach to train a deep neural network in a layer-wise fashion. In contrast to E2E learning, whereby all layers of a network are trained simultaneously, CL trains a network one layer at a time, freezing the previously trained layers. The aim is to circumvent the vanishing gradient problem by ensuring that the output is always adjacent to the layer being trained. However, in this paper we show that such training also leads to improvement in several transfer learning tasks. Inspired by *Belilovsky et al.* (Belilovsky et al., 2019), when training the n^{th} layer of the cascade convolutional filters, we add a randomly initialised auxiliary classifier (AC), consisting of k convolutional layers and f fully connected layers. AC_S, denoting an AC at the source, has been shown to improve performance (Belilovsky et al., 2019). We set k and f to be small because our goal is to transfer features from pretrained models and not make the target domain networks too complicated. Note, we add an AC to the E2E framework as well, ensuring identical architectures for a fair comparison across learning paradigms for transfer as demonstrated in Figure 1. The complete network architecture is shown in Table S2. Code is available: https://github.com/FrankWJW/cascade_transfer_learning_medical

2.2. Transfer with Cascade Learning

When transfer via cascade learning (TCL) is applied at the n^{th} layer, the network from the source domain up until layer n (including layer n though excluding AC_S), is copied to the

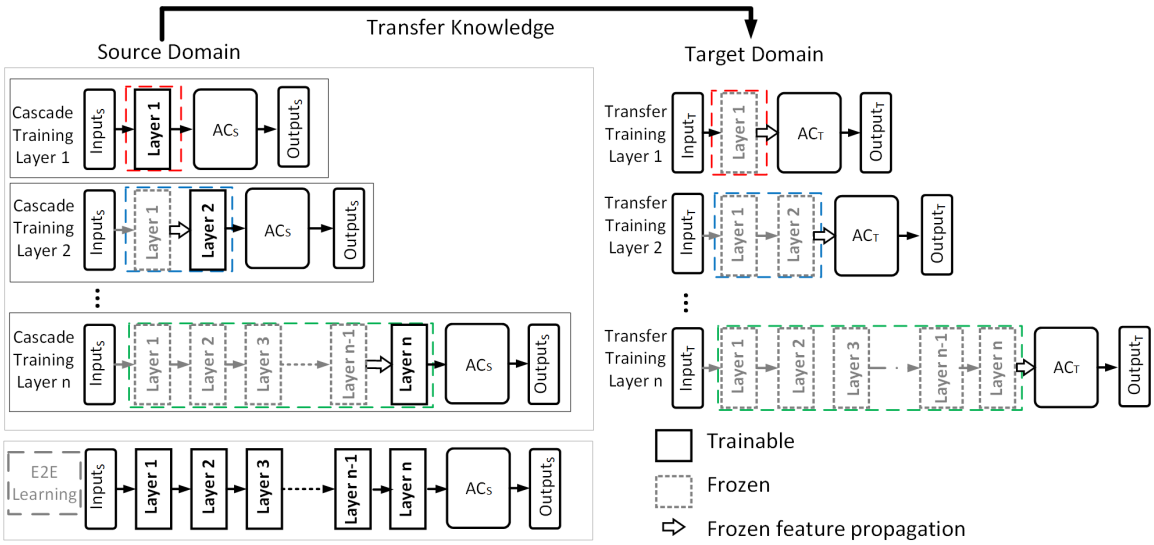


Figure 1: Schematic illustration of cascade learning (CL), transfer cascade learning (TCL) and end-to-end (E2E) learning. AC_S and AC_T refer to the auxiliary classifiers in the source and target domains, respectively. Note, AC_S is also included in the E2E framework ensuring transfer comparison across learning paradigms.

target domain. This is illustrated by bounding boxes in Figure 1. The auxiliary classifier in the target domain, denoted AC_T, is randomly initialised and trained on the target data. We compare the performance of transferring each CL versus E2E individual layers, observing significant differences in the features learned and improved performance with TCL, in addition to other advantages. Considering time complexity (wall clock) within limited data regimes (i.e. small to medium sized datasets) where transfer learning is beneficial, the difference in training time between CL and end to end learning is negligible.

2.3. Source Domain Datasets

We use natural images from the ImageNet dataset (Deng et al., 2009) as the source for all of our experiments. ImageNet pre-trained models are widely used as the source of transfer (Ke et al., 2021; Shin et al., 2016; Tajbakhsh et al., 2017; Wang et al., 2017).

In this paper, we are comparing learning paradigms which would require training ImageNet from scratch. As this is infeasible due to limited computing resources and time, particularly for CL paradigms. Instead we consider a subset of ImageNet, which we refer to as ImageNet23. This subset includes $29,900$ natural images, obtained from the 23 classes that overlap with the CIFAR-100 dataset (Krizhevsky, 2009). This choice is rather arbitrary and done with the purpose of keeping the feature extractor simple. Searching for a good source problem, i.e. classes in the source domain that could offer better feature, is an open though computationally intractable problem.

2.4. Target Domain Datasets

We conduct classification experiments on five target datasets considering a range of medical imaging tasks and target dataset sizes. We consider two chest X-ray datasets, one of which is imaging of Covid-19 patients, as well as gastrointestinal disease detection via endoscopy images, ductal carcinoma from whole slide images, and a skin lesion detection dataset via dermatoscopic images. Our localisation experiments are conducted on a separate dataset as this is the only one that contains bounding box labels required for the task. The datasets used in this paper are all publicly available. We summarize some of their attributes in the following paragraphs.

Kvasir (Pogorelov et al., 2017) is a dataset containing endoscopy images of gastrointestinal tracts. Each endoscopy image is labelled by one or more medical experts from Vestre Viken Health Trust (VV) and Cancer Registry in Norway. The data consists of 4000 images with 8 classes showing the anatomical landmarks (Pogorelov et al., 2017). Each class contains 500 images with different resolutions: 720×576 to 1920×1072. We resize images to 256×256 by downsampling.

The first iteration of the **BIMCV COVID-19+** (Vayá et al., 2020) dataset contains 1380 chest X-rays as well as associated medical reports and metadata from the Valencian Region Medical ImageBank (BIMCV). Up to September 2021, the dataset contained $13,615$ samples from $4,896$ patients. In order to select data for this paper, we perform two-steps: first, we select samples where the modality is chest X-ray; second, we select samples with the associated finding containing the keywords Covid-19 or normal. After the selection process, we obtain in total 3705 samples, which we store as uncompressed gray-scale *.png* file, in 16 bit format. We conduct an additional pre-processing step using DICOM *WindowCenter* and *WindoWidth*, converting to *Monochrome 2* photometric interpretation, resizing the images to 224×224 resolution (Vayá et al., 2020) followed by data augmentation. To be consistent with the pre-trained model, the images are broadcast to 3 channels.

The Invasive Ductal Carcinoma **(IDC)** (Cruz-Roa et al., 2014) dataset was collected from 162 patients diagnosed with IDC. 113 slides of Whole Slide Image (WSI) were selected for training and 49 slides were held out for testing. Each WSI is sliced into 50×50 image patches, with corresponding ground truth labelled by an expert pathologist. This process yields over $21,000$ labelled image patches. The label distribution of this dataset indicate a strong class-imbalance, hence weighted sampling is used to keep mini-batches class-balanced during training. To be consistent with pre-trained models, the image was up-sampled into 224×224 resolution.

The **HAM10000** (Tschandl et al., 2018) dataset is a large collection of multi-source skin lesion images. The original data consists of $10,015$ samples as a training set. There are in total 7 classes corresponding to different categories of pigmented lesions. Original image from the dataset has resolution 600×450 in *.jpg* format. The experimental setup and handling of class-imbalance is the same with the IDC dataset.

CheXpert (Irvin et al., 2019) is a large dataset containing $224,316$ chest X-ray images from $65,240$ patients. Each data instance has multiple binary labels that represent positive or negative observations for 14 types of disease. Multiple positive observations can be labelled in a chest X-ray image. Even though the dataset introduces uncertainty in the labels, in our experiment, we focus on transferring features using CL and ignore uncertain

labels by mapping them into negative cases. The data has been resized to 224×224 and broadcast to 3 channels.

For our localisation experiments in Figure 6, we use the **ChestX-ray8** (Wang et al., 2017) as it contains board-certified radiologist bounding box labels. While ChestX-ray8 contains $108,948$ frontal view chest X-ray images from $32,717$ patients (Wang et al., 2017), we use the 983 images which have been labelled with bounding boxes in order to evaluate our proposed method.

2.5. Feature Visualisation via Grad-CAM

Feature visualisation is particularly important in the medical context to add interpretability to the deep neural network predictions (Arias-Londono et al., 2020; Irvin et al., 2019; Reyes et al., 2020; Simonyan et al., 2014; Wang et al., 2017). One of the previous work, **saliency maps** (Simonyan et al., 2014) is one way of interpreting the effect of each pixel given an input image I on the final prediction. This is done by taking the gradient of the class score (S_c) with respect to the input image itself as follows:

$$w = \frac{\partial S_c}{\partial I} \tag{1}$$

The result will give us an activation map of the degree to which a pixel contributed to that class score. This gives us insight into what the network is focusing on with respect to the input image for each particular class prediction.

The **Grad-CAM** (Selvaraju et al., 2020) method generates a heat-map of the input pixels, telling us where the model is looking at to make a particular prediction. Grad-CAM considers how a change in a particular location i, j, in the activation map A^k, creates a change in the class activation y_c by computing this gradient (Equation 2). This is accumulated by summing the values over the entire activation map indexed by k to give α_k^c. The scalar α_k^c represents *neuron importance* for the k^{th} feature map and class c. Finally, $L_{\text{Grad}-\text{CAM}}$ computed as Equation 3, where Z denotes the total number of pixels in the feature map. Equation 3 accumulates the neuron importance over all the activation maps, followed by the ReLU non-linearity to remove the negative components. $\alpha_k^c < 0$ implies that a change in A^k will decrease prediction score y^c, which should be avoided as those feature maps that improve the prediction are of interest (Selvaraju et al., 2020), hence the ReLU.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

$$L_{\text{Grad}-\text{CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{3}$$

2.5.1. Granulometry Analysis

Features extracted by cascade and end-to-end trained models differ in their distributions. We postulate that cascade learning extracts **coarse** features in its early layers and progressively **finer** ones in later layers owing to each layer being trained having limited flexibility

of function fitting. This, we believe is one reason for early layers of cascade trained models offering better transfer to a different domain because these layers have not learnt fine details of the source domain. We use an image processing technique known as *granumoletry* (Dougherty et al., 1989) to quantify this difference applied to the sensitivity maps between images and corresponding class predictions. Granulometry is a quantitative measure of how contiguous regions in an image space are, with coarse images carrying higher values than those in which features are distributed as large numbers of small patches. The morphological image processing operations leading to this analysis are shown in Algorithm 1. We also give a synthetic example illustrating this in Appendix Figure S1.

In our implementation of granulometry, we take activated regions with the top 25% brightness to do binarisation, filtering out regions with negligible brightness. From the distribution of different sized regions, we take the number of median regions as the output of the granulometry algorithm, avoiding the case where all active regions have low brightness. We also include count of connected areas to improve the granulometry measure. The count of connected areas is calculated as the area of connected regions divided by the number of fix sized elements. For example, if the area of connected regions is fixed for a mask, the increasing number of small connected grains will result in small granulometry score as the activated regions are scattered instead of concentrated. Finally, from each layer we get a normalised granulometry score. We use *scipy.ndimage* library to compute morphological image processing operation. We are providing code for an illustrative problem. Please refer to section 2.1 for the GitHub link.

2.5.2. Robustness to Noise and Confidence Calibration

Two other criteria by which medical applications of computer vision should be judged are robustness to noise and calibration. A trained model, when deployed, should be able to cope with instrument noise. Noise could be of two types: systematic effects such as calibration errors or instruments made by different manufacturers, or random noise arising in acquisition. Here, we address random additive noise in the measurement process and quantify how much the different methods considered degrade at increasing levels of signal to noise ratio. Addressing systematic variations (which are also known as covariate shift) is beyond the scope of the present work.

The second issue we consider is calibration. Confidence calibration is another important aspect for the successful deployment of deep learning systems, particularly in automatic health care (Guo et al., 2017). A well calibrated model will inform human experts to take over the final decision when the confidence of diagnosis is low. Recent analysis shows that modern deep neural networks are not well calibrated (Guo et al., 2017), meaning their prediction probability estimates are not representative of the true correctness likelihood of the labels. In this paper, we measure *Expected Calibration Error (ECE)* (Guo et al., 2017) to quantify the model uncertainty:

$$\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) = p, \quad \forall p \in [0, 1] \tag{4}$$

$\hat{Y}$ is the predicted label and Y is a random variable based on ground truth; $\hat{P}$ is the confidence score (represented by softmax probability) and p is another random variable representing a group of confidence score have value p. Then ECE is computed from Equation 4 by using a binning method to split data into equal-sized bins. Data belonging to

each bin has the same confidence level, then the absolute difference between the accuracy and confidence within the group of data in the bin B_m is measured as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}\left(B_m\right) - \mathrm{conf}\left(B_m\right)\right| \tag{5}$$

where m is the bin index and n is number of samples.

Algorithm 1 Pseudocode of Granulometry Score Computation

Data: A: A saliency map of an image with size $m \times n$
Result: GS: The granulometry score;
Initialisation:
$B(u) \leftarrow$ A structuring element with size u, e.g., an ellipse
$AS \leftarrow$ The list of grain (connected areas) size
$M \leftarrow$ Binary mask of A. Each element of the mask is "True" or "False"
$C \leftarrow$ A list of grain counts.
$M \leftarrow$ A $< 75\%$ quantile of A. ▷ This step filters out regions having low brightness in the saliency map.
$AS \leftarrow [1, \mathrm{int}(0.25 \times \max(m,n)/20), \mathrm{int}(0.75 \times \max(m,n)/20), \mathrm{int}(\max(m,n)/20)]$[1]

for s *in* AS **do**
  $newA \leftarrow$morphological_opening$(M, B(s))$.
  $C \leftarrow$ append the counts of connected areas in $newA$.
end
GS $\leftarrow$ the median of C.

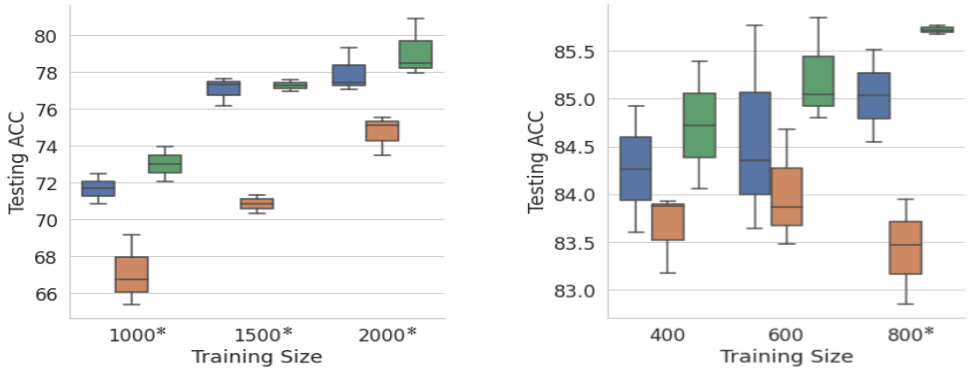
3. Results

3.1. An Comparison of Varied Learning Approaches for Transfer

We have performed numerous experiments comparing transfer with CL, transfer with E2E learning, as well as learning directly on the data (full range of experiments documented in the Appendix). In general, we conclude that TCL outperforms TE2E particularly on the smaller data regimes, reconfirming previously published work (Du et al., 2019) and also reconfirming that smaller models tend to transfer better (Ke et al., 2021).

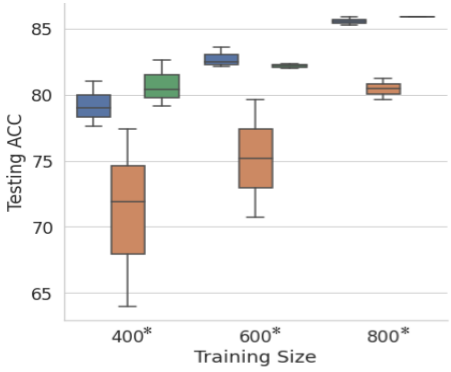
The results in Figure 2 summarise the performance of TCL versus TE2E, as well as the combined approach on five distinct medical datasets. Note, the results in this figure have been tuned for optimal hyper-parameter selection over all experiments. Figure 2 demonstrates the superior performance of TCL against TE2E on all of the medical imaging tasks with statistically significant cases ($p \leq 0.05$, two-sided pair sample t-test) denoted by $*$ on the x-axes. Furthermore, due to errors made by TCL and TE2E approaches being different, model combination achieves better classification accuracies in all cases. We empirically show that a combined approach can improve the performance in all cases.

1. Numbers are chosen based on trial and error.

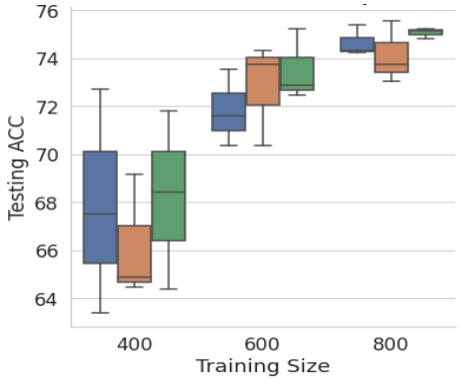


(a) HAM10000

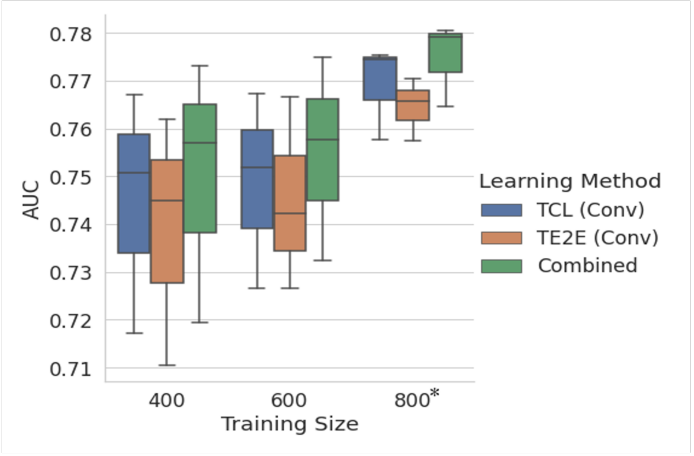
(b) IDC



(c) Kvasir



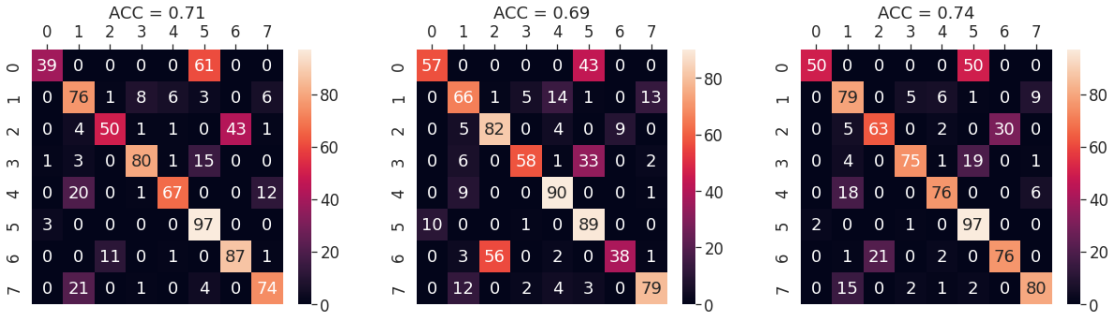
(d) BIMCV



(e) CheXpert

Figure 2: Performance comparison on five medical datasets. The combined approach consistently improves performance. We report statistically significant differences between TCL and TE2E via two-sided paired sample t-test ($p \leq 0.05$), with significant cases labeled by $*$ on the x-axes.

Interestingly, the error analysis of TCL versus TE2E Figure 3 reveals different errors made by the learning paradigms ((a) vs. (b)), and therefore likely different feature representation learning. We use these results to try a combined transfer approach, averaging the predicted outputs of both models. Full results of combining approach are shown in Figure 2. We also demonstrate via Grad-CAM saliency maps and granulometry measures that the features learned are indeed different, and provide quantitative and qualitative evidence that TCL features are more localised, providing better descriptors for interpretability and prediction. We found TCL outperforms current self-supervised transfer learning framework with direct comparison to the reported result in (Truong et al., 2021). The result shows in Table S1. We provide an evidence that even using a network architecture with low parameter complexity and train with lower volume source data, CL could still achieve large improvement even comparing against complex framework trained via full scale ImageNet dataset, in the medical domain.



(a) TCL (Layer 3) (b) TE2E (Layer 3) (c) Combined (Layer 3)

Figure 3: Confusion matrices of **(a)** TCL and **(b)** TE2E learning methods on the Kvasir (Pogorelov et al., 2017) dataset. These results demonstrate that different errors are made, motivating a combined approach **(c)**. The combined approach is the average of the predicted softmax outputs of both models.

Table 1 presents the TCL results on the entire CheXpert dataset for comparison to ResNet-50 (Raghu et al., 2019). We note that TCL outperforms training from scratch (ResNet-50) on this large dataset in all five disease categories, with significant improvements in three out of the five classes, motivating the use of TCL even in large data regimes.

3.2. Feature Representation of TCL

Next, we ask the question *are the features learned by TCL and TE2E different? Further, which are more localised to the target area of interest?* We begin by plotting the Grad-CAM saliency maps to probe the feature maps.

We visualise the Grad-CAM saliency maps for TCL versus TE2E on the ChestX-ray8 dataset that contains radiologist bounding box labels (red box in Figure 4), demonstrating the TCL features found at the intermediate layers to be more closely aligned. Our large scale quantitative experiments, described in more detail later in this section, demonstrate the repeated superior localisation accuracy of TCL in comparison to all other methods.

Table 1: Performance of TCL in comparison to E2E trained source models on the CheXpert dataset. Note, the E2E results (TResNet-50 and ResNet-50 are taken from the literature (Raghu et al., 2019)). Results are reported in AUC-ROC. (*) indicates a statistically significant difference via two-sided pair sample t-test ($p \leq 0.05$).

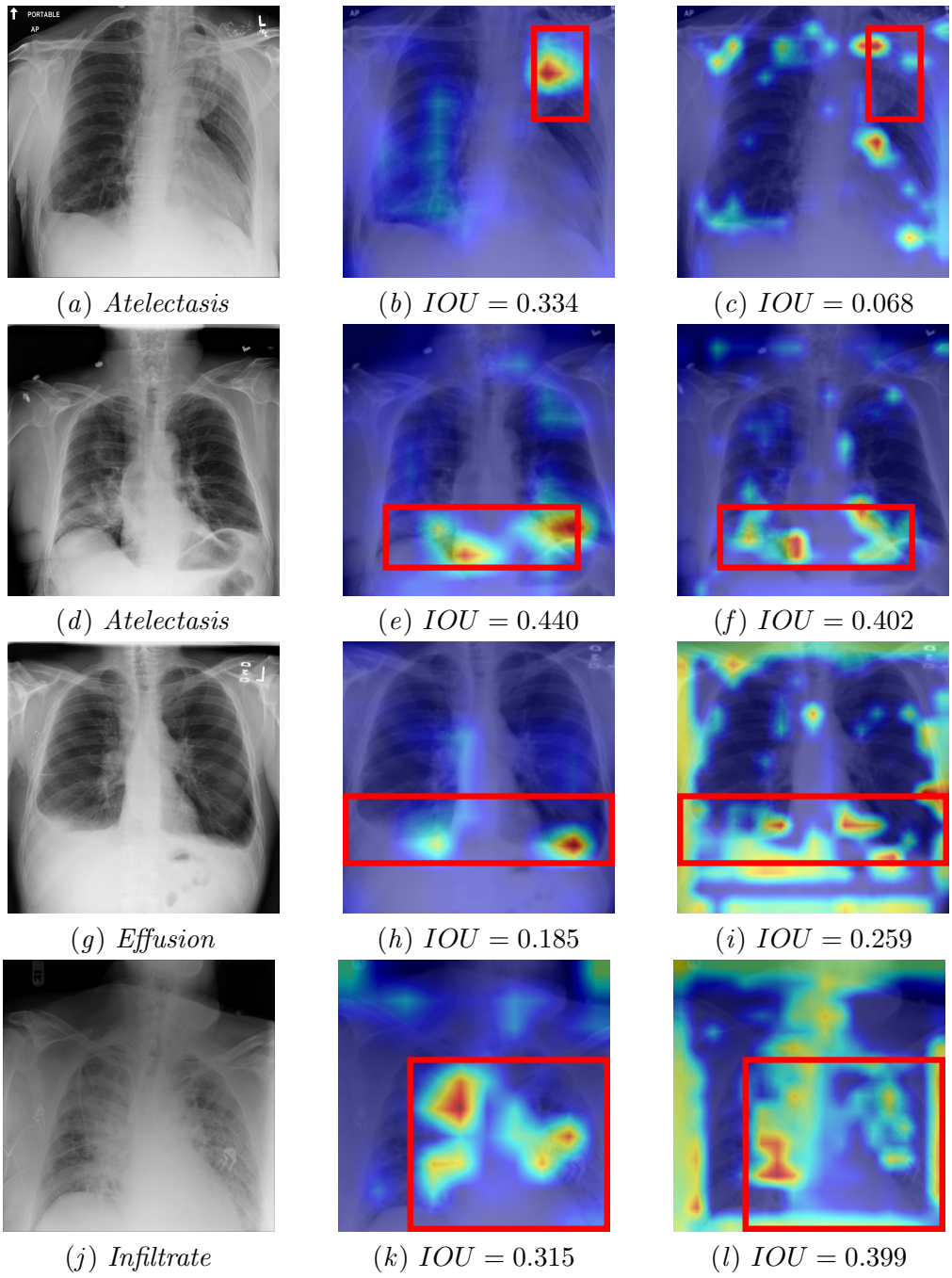
Disease	*TCL*	*TResNet-50*	*ResNet-50*
Atelectasis	**79.80**±0.93	79.76±0.47	79.52±0.31
Cardiomegaly	**78.68**±0.93 (∗)	74.93±1.41	75.23±0.35
Consolidation	**90.16**±0.68 (∗)	84.42±0.65	85.49±1.32
Edema	**89.87**±0.11	88.89±1.66	88.34±1.17
Pleural Effusion	**91.33**±0.13 (∗)	88.07±1.23	88.70±0.13

Further, we visualise the Grad-CAM saliency maps over several layers and across all of the methods in Figures S2 and S3, demonstrating more consistent localised features with TCL. Pleural Effusion is a disease type with excess liquid build up at the pleura region (Hooper et al., 2010). We observe in Figure S2 that the CL-trained network achieves improved localisation at nearly every layer as observed by the red patches near the bottom of the lungs, while the E2E network achieves the correct location merely at last convolutional layer.

Figure 4(b), 4(e), 4(h) and 4(k) shows that saliency map generated via TCL is having coarse feature and accurately localised to target ROI. A reasonable question to ask is whether this observation is hold for every images, as oppose to cherry-picking the result that have best the visualisation.

Granulometry analysis (Dougherty et al., 1989) on the generated saliency maps quantitatively demonstrate the coarse-to-fine feature representation of TCL versus TE2E. The higher granulometry represents the feature activation (indicated as the irregular red patch in Figure 4). The results in Figure 5 demonstrate that TCL generates saliency maps that have higher granulometry, indicating the feature activations are coarser, at early layers and finer at later layers. TE2E has very evenly distributed granulometry across the layers. These results further strengthen the argument for CL in transfer as we demonstrate that early layers in the network are learning coarser features while later layers are learning more fine grained features.

We are interested to know if features learned by TCL are able to accurately identify target regions of interest consistent with human expert labels. In order to address this question, we quantitatively compare our saliency map results with pixel-level annotations done by medical experts available only in the *Chest-Xray8* dataset. We measure the *Intercept Over Union* (IOU) between our saliency maps and the region within the bounding box labels. We assume $IOU > 0.1$ to be correctly localised, then measure the localisation accuracy (i.e. number of images that are correctly localised) over 983 instances. Figure 6 shows the overall localisation accuracy comparisons over the five learning methods over the intermediate layers. There is a large improvement in localisation accuracy with TCL over the other learning methods, with TCL constantly obtaining higher localisation accuracy



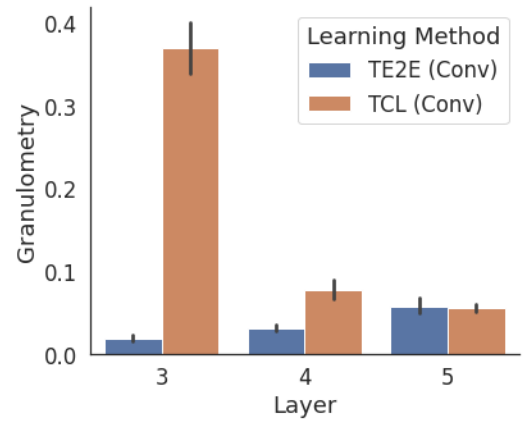
(*a*) *Atelectasis* (*b*) $IOU = 0.334$ (*c*) $IOU = 0.068$

(*d*) *Atelectasis* (*e*) $IOU = 0.440$ (*f*) $IOU = 0.402$

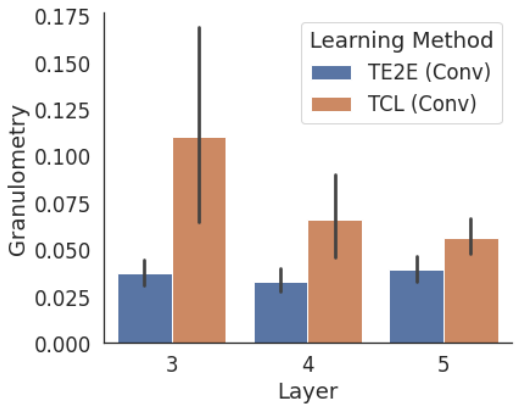
(*g*) *Effusion* (*h*) $IOU = 0.185$ (*i*) $IOU = 0.259$

(*j*) *Infiltrate* (*k*) $IOU = 0.315$ (*l*) $IOU = 0.399$

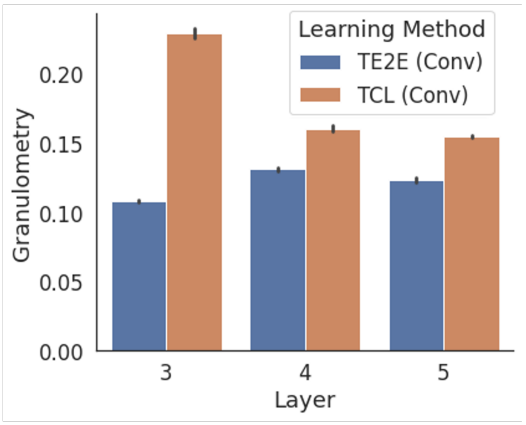
Figure 4: Grad-CAM generated via TCL intermediate layer in comparison to the same layer from TE2E. Red rectangle denotes ground truth bounding box annotated by the clinician. The TCL features learned are localised to the target region of interest. Left column: Original image and its finding label; Middle: **TCL**; Right: **TE2E**. Both TCL and TE2E have same architecture, feature maps are taken at third layer.



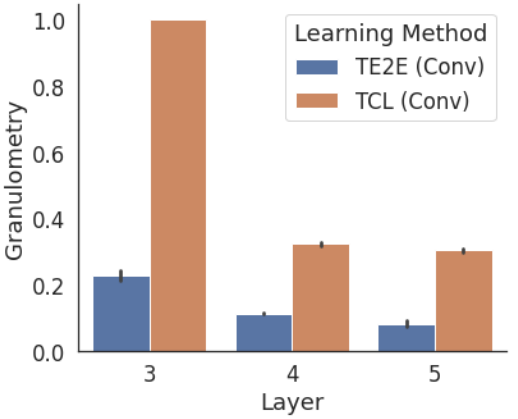
(*a*) Dataset = IDC



(*b*) Dataset = Kvasir



(*c*) Dataset = HAM10000



(*d*) Dataset = BIMCV

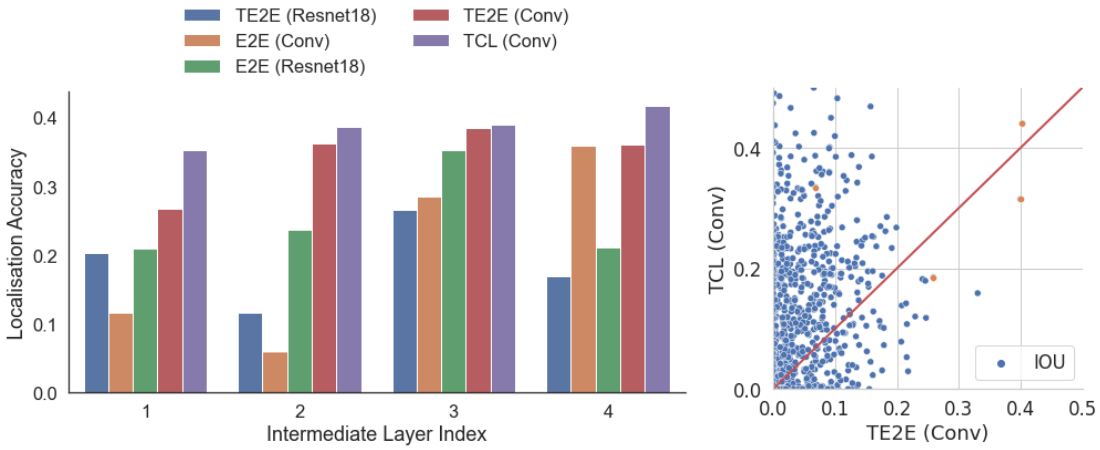
Figure 5: Granulometry measure comparing TCL and TE2E learning methods on different layers. Result demonstrating on four distinct medical problems.

over all layers of the network, greatly outperforming TE2E. This result is significant for interpretability, demonstrating that TCL is more often making predictions using the same regions of the images that medical experts identify.

3.3. Noise Robustness and Confidence Calibration

In Figure 7(*a*), we seek to answer the question *is TCL more robust to noise?* by adding additive white Gaussian noise to the target domain. We demonstrate that a small amount of noise results in TE2E performance dropping severely, although this level of noise is not visually perceivable.

In confidence calibration analysis, we use the *ECE* to quantify which learning method yields better calibration. Note, ECE is an error measure and therefore lower values are desired. In the clinical setting, lower ECE could indicate the model which produces a confidence score that is less likely to be overconfident (e.g. false positive prediction with



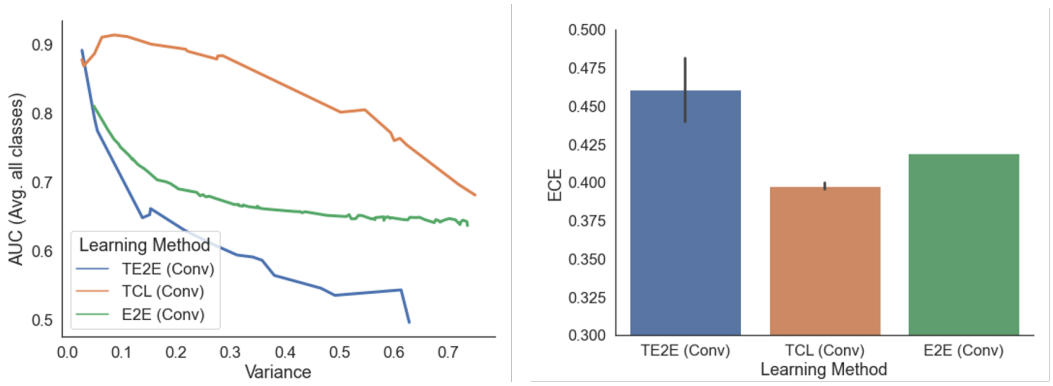
(*a*) Localisation Accuracy (*b*) IOU

Figure 6: TCL consistently results in better localisation, meaning the features learned are more closely aligned with the labelled regions of interest. a) The localisation accuracy, measured as the IOU between the saliency maps and bounding boxes, over varied learning method layers. The error bar on average IOU across different trials did not appear due to small variation. However, there is large variation in IOU across images. b) Scattering plot of IOU between the manual annotation and saliency maps. Results demonstrated using 983 chest X-ray images with manual annotation from the Chest-Xray8 dataset. Orange dots represent example images used in Figure 4.

a confidence probability 0.99). In such case, a well-calibrated model should return a lower confidence value to indicate that their aware they are likely making incorrect predictions so that the clinician can take over the decision-making process. Figure 7(*b*) shows that TCL tends to have lower ECE than E2E and TE2E learning method, implying the predictions are more representative of the true likelihoods of the labels.

4. Conclusions

In this paper, we demonstrate for the first time the superior feature representations learned using CL, considering five medical imaging classification tasks and one localisation task. We show that a layer-wise learning strategy has many advantages in comparison to traditional learning, particularly as the way the feature representation learning is achieved varies. This difference not only offers improved transfer performance, it also offers many other potential advantages. By exploring the types of errors made and the types of features learned, we discover that TCL offers improvements in terms of the localisation of features, robustness to noise, as well as improved calibration further supporting the hypothesis that learning in a layer wise fashion is a superior strategy for medical image classification. Overall, we demonstrate empirically that TCL is a superior learning method for deployment into real



(*a*) Performance over adding white Gaussian noise with changing variance (x-axis).

(*b*) Model calibration errors, measured using ECE.

Figure 7: The results demonstrate that TCL tend to be robust to noise and better calibration in comparison to E2E and TE2E. The experiments are conducted on the HAM10000 dataset.

clinical environments in terms of classification performance, tractable activations, noise robustness and confidence calibration.

References

John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Convolutional neural networks for mammography mass lesion classification. In *the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 797–800. IEEE, 2015.

Julian D. Arias-Londono, Jorge A. Gomez-Garcia, Laureano Moro-Velazquez, and Juan I. Godino-Llorente. Artificial Intelligence applied to chest X-Ray images for the automatic detection of COVID-19. A thoughtful evaluation approach. *IEEE Access*, 8, 2020. doi: 10.1109/ACCESS.2020.3044858.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021.

Yaniv Bar, Idit Diamant, Lior Wolf, and Hayit Greenspan. Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging: Computer-Aided Diagnosis*, 2015.

Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Greedy layerwise learning can scale to Imagenet. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 583–593. PMLR, 2019.

Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

Hao Chen, Dong Ni, Jing Qin, Shengli Li, Xin Yang, Tianfu Wang, and Pheng Ann Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1627–1636, 2015.

Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Medical Imaging 2014: Digital Pathology*, volume 9041, pages 1–15. International Society for Optics and Photonics, SPIE, 2014. doi: 10.1117/12.2043872.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

C Di Ruberto, A Dempster, S Khan, and B Jarra. Automatic thresholding of infected blood images using granulometry and regional extrema. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 441–444 vol.3, 2000. doi: 10.1109/ICPR.2000.903579.

Edward R Dougherty, Eugene J Kraus, and Jeff B. Pelz. Image segmentation by local morphological granulometries. In *Proceedings of the 12th Canadian Symposium on Remote Sensing Geoscience and Remote Sensing Symposium*, volume 3, pages 1220–1223. IEEE, 1989.

Xin Du, Katayoun Farrahi, and Mahesan Niranjan. Transfer learning across human activities using a cascade neural network architecture. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC)*, pages 35–44. ACM, 2019.

Xin Du, Katayoun Farrahi, and Mahesan Niranjan. Information Bottleneck Theory Based Exploration of Cascade Learning. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101360.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

Scott E. Fahlman and Christian Lebiere. *The Cascade-Correlation Learning Architecture*, page 524–532. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1558601007.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6391–6400, 2019.

P. Gravel, G. Beaudoin, and J.A. De Guise. A method for modeling noise in medical images. *IEEE Transactions on Medical Imaging*, 23(10):1221–1232, 2004. doi: 10.1109/TMI.2004.832656.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On Calibration of Modern Neural Networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1321–1330. PMLR, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Clare Hooper, Y C Gary Lee, and Nick Maskell. Investigation of a unilateral pleural effusion in adults: British Thoracic Society pleural disease guideline 2010. *Thorax*, 65(Suppl 2): ii4—-ii17, 2010. ISSN 0040-6376. doi: 10.1136/thx.2010.136978.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Ciurea-Ilcus, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Visakan Kadirkamanathan and Mahesan Niranjan. A function estimation approach to sequential learning with neural networks. *Neural Computation*, 5(6):954–975, 1993.

Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y. Ng, and Pranav Rajpurkar. Chextransfer: Performance and parameter efficiency of imagenet models for chest x-ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 116–124, 2021. doi: 10.1145/3450439.3451867.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

Sindy Löwe, Peter O'Connor, and Bastiaan Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations. *Advances in Neural Information Processing Systems*, 32, 2019.

Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, et al. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology*, 294(2):421–431, 2020. doi: 10.1148/radiol.2019191293.

Jan Margeta, Antonio Criminisi, R Cabrera Lozoya, et al. Fine-tuned convolutional neural nets for cardiac mri acquisition plane recognition. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 5(5):339–349, 2017.

Enrique S. Marquez, Jonathon S. Hare, and Mahesan Niranjan. Deep cascade learning. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):1–11, 2018.

Arild Nøkland and Lars Hiller Eidnes. Training neural networks with local error signals. In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

John Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225, 1991.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, et al. KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings - ACM Multimedia System*. Association for Computing Machinery, 2017.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 3347–3357. Curran Associates, Inc., 2019.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020.

Stephen Roberts and Lionel Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6(2):270–284, 1994.

Thomas Schlegl, Joachim Ofner, and Georg Langs. Unsupervised pre-training across image domains improves lung tissue classification. In *International MICCAI Workshop on Medical Computer Vision*, pages 82–93. Springer, 2014.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. ISSN 15731405. doi: 10.1007/s11263-019-01228-7.

Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1099, 2015.

Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5): 1285–1298, 2016. doi: 10.1109/TMI.2016.2528162.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings (2014)*, 2014.

Patrick S Stumpf, Xin Du, Haruka Imanishi, Yuya Kunisaki, Yuichiro Semba, et al. Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell rna sequencing. *Communications Biology*, 3(1):1–11, 2020.

Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2017. ISSN 23318422.

Tuan Truong, Sadegh Mohammadi, and Matthias Lenga. How transferable are self-supervised features in medical image classification tasks? In *Machine Learning for Health*, pages 54–74. PMLR, 2021.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161.

Bram Van Ginneken, Arnaud AA Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 286–289. IEEE, 2015.

Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, et al. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. *arXiv*, June 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, et al. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Yulin Wang, Zanlin Ni, Shiji Song, Le Yang, and Gao Huang. Revisiting locally supervised learning: an alternative to end-to-end training. In *International Conference on Learning Representations (ICLR)*, 2021.

Tianyang Zhang, Jun Cheng, Huazhu Fu, Zaiwang Gu, Yuting Xiao, et al. Noise Adaptation Generative Adversarial Network for Medical Image Analysis. *IEEE Transactions on Medical Imaging*, 39(4):1149–1159, 2020. doi: 10.1109/TMI.2019.2944488.

Appendix A.

Performance Comparison with Self-supervised Features in Transfer

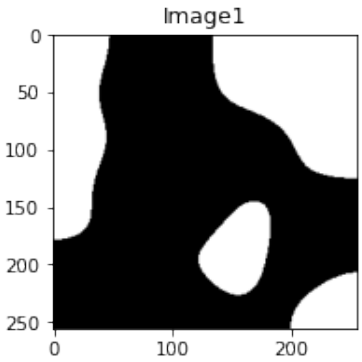
We compare our method against two self-supervised methods including the current state-of-the-art (Truong et al., 2021). We conduct experiment on *NIH Chest X-ray dataset* (Majkowska et al., 2020), which released two subsets of chest X-ray images. For each subset have 2414 and 1962 chest X-ray images, respectively. There are four medical conditions annotated by radiologist (pneumothorax, nodule or mass, opacity, and fracture). We include the class *Normal* indicating no medical conditions with respect to the four class labels (Truong et al., 2021). Table S1 shows a comparison between TCL and self-supervised methods DINO and SimCLR. The result are averaged over 3 runs with 5 fold cross validation for each run.

Table S1: The mean AUC is obtained across 5 folds and 3 training set partitions. Note, the results for *DINO* and *SimCLR* are taken from the literature (Truong et al., 2021).

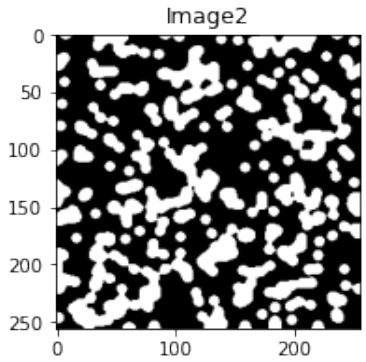
	Training Size	TCL	DINO	SimCLR
Without Fine Tune	50	0.6614±0.0018	**0.6831±0.0233**	0.6273±0.0130
	200	0.7134±0.0065	**0.7373±0.0112**	0.6645±0.00067
	2414	**0.7589±0.0018**	0.7438±0.0228	0.6983±0.0231
With Fine Tune	50	**0.6529±0.006**	0.6348±0.0286	0.6227±0.0309
	200	0.684±0.0024	0.6652±0.0114	**0.7228±0.0287**
	2414	**0.7643±0.0054**	0.7404±0.0240	0.7358±0.0295

Additional Explanation of Granulometry

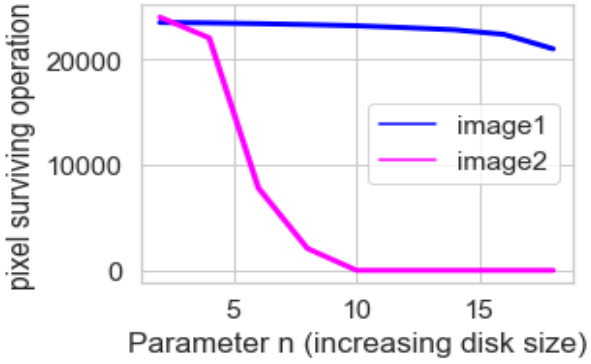
An important observation we make about cascade learning is that features extracted during layer-wise training tend to be coarse in early layers. This is because when training one layer at a time, the network is constrained to have limited flexibility. As we go to deeper and deeper layers, finer details specific to the domain in which the network is trained are extracted. Granulometry Dougherty et al. (1989) is an image processing technique that can help quantify this effect and has been used in applications such as foreground-background segmentation of blood cell images (Di Ruberto et al., 2000). In Figure S1, we give a simple illustration to how coarse and fine distribution of features in the space of GradCam maps differ in this measure.



(*a*) Example image (coarser)



(*b*) Example image (finer)



(*c*) Pixel survived against disk size

Figure S1: Granulometry measure on two synthetic images. The two black-and-white synthetic images have the same proportion of white space. One has large continuous regions (coarse distribution) and in the other features are distributed as finer lumps. Granulometry is computed by moving around the space disks of varying sizes (defined by parameter n) until they touch black pixels. The surviving pixel is defined by number of positive pixels (white color blobs in the image) are remain positive after the morphological opening operation. The coarser image has large continuous region therefore the large size disk will not touch the black pixels if move the disk within the region.

Implementation Details

The Simple CNN structure used in the experiments labelled *Conv* can be found in Table S2.

Table S2: Simple CNN architecture used in the experiment.

	layer name	output size	11-layer
Feature	conv1_f	114 x 114	[3x3, 256] x 1
	conv2_f	59 x 59	[3x3, 256] x 1
	conv3_f	31 x 31	[3x3, 256] x 1
	conv4_f	17 x 17	[3x3, 256] x 1
	conv5_f	10 x 10	[3x3, 256] x 1
	conv6_f	7 x 7	[3x3, 256] x 1
	conv7_f	5 x 5	[3x3, 256] x 1
	conv8_f_x	4 x 4	[3x3, 128] x 4
Classifier	conv1_c	4 x 4	[3x3, 128] x 1
	flatten	1 x 1	N/A
	fc_1	N/A	256
	fc_2	N/A	256
	fc_3	C	256

All the networks are trained with standard data augmentation which includes random cropping and flipping, random rotations and image normalization. For hyper-parameter tuning in Figure 2, we search for optimal learning rate using and batch size using the following setting: Initial learning rate sampled from 10^{-5} to 10^{-1} on log space; mini-batch size sample from $(8, 16, 32, 64)$. We use Adam as optimization algorithm. We select the best performing model between TCL and TE2E on the validation set. We use this best performing model to make inference on the test data. For the combined approach, we average their combined softmax prediction on the test data and report this result. We use the hyper-parameter tuning library *Ray tune* (Liaw et al., 2018) for systematic hyper-parameter tuning. All experiment are running on a single machine with 4 *RTX 2080* GPU and use *Pytorch* (Paszke et al., 2019) machine learning framework.

More results on GradCAM analysis

Figure S2 shows GradCAM visualisation on both CL and E2E training on target domain data (chest X-ray) over various intermediate layer.

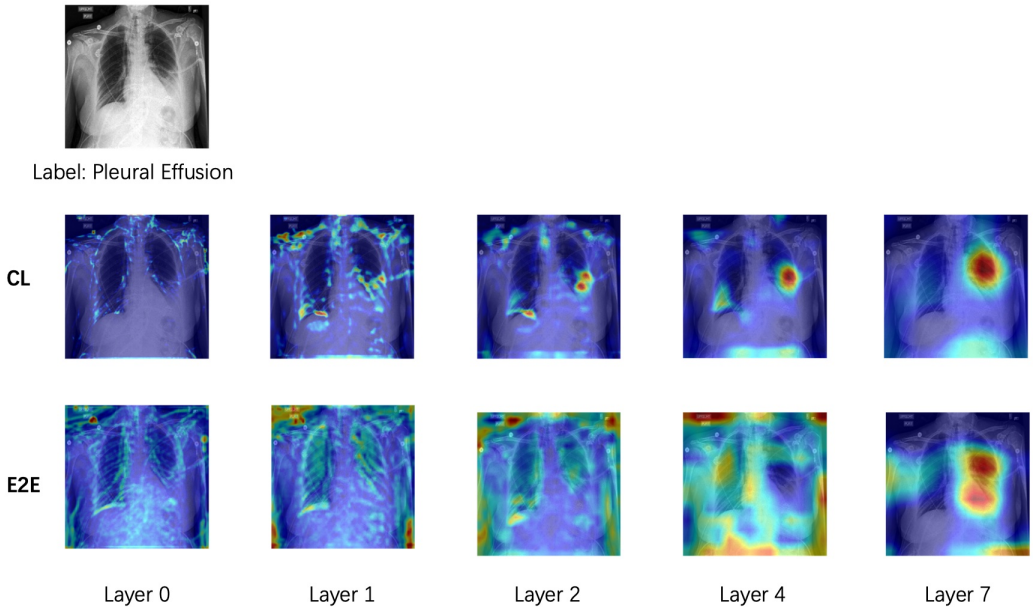


Figure S2: Grad-CAM saliency map at different layers. Top: Results on a cascade trained network. Bottom: Results on E2E training. In comparison to E2E training, CL achieves better localisation of the target for every layers. Result demonstrated using CheXpert dataset.

Figure S3 shows GradCAM visualisation on five learning method over three medical datasets with different modality.

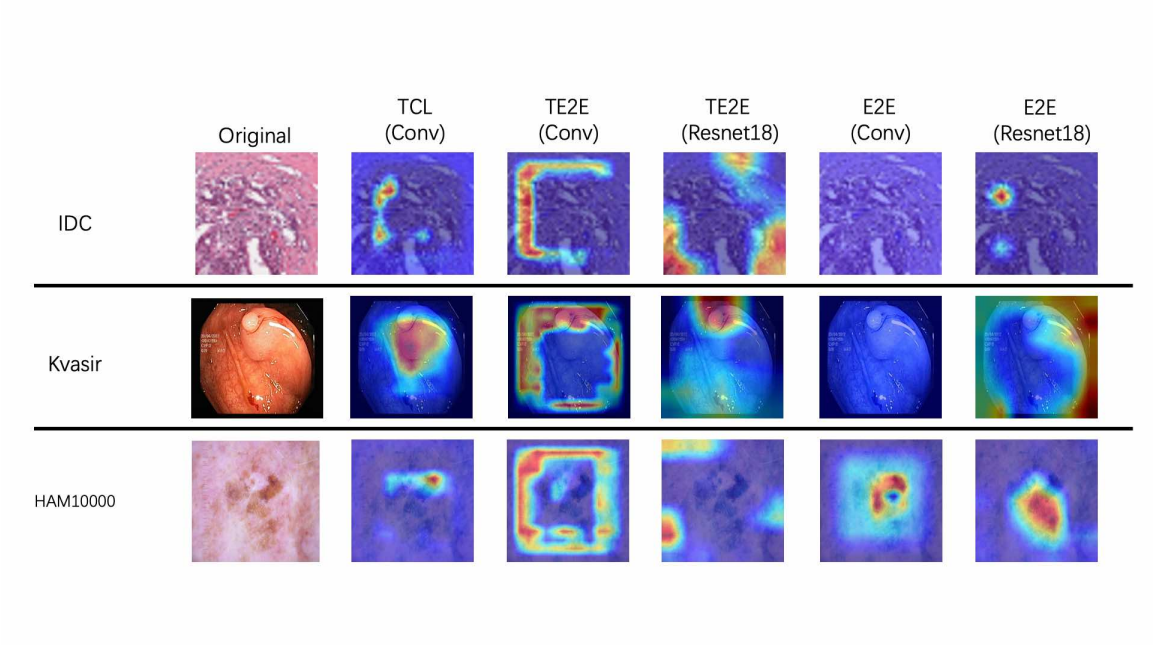


Figure S3: Grad-CAM visualisations for all five learning methods at last convolutional layers across three distinct medical problems. From left to right, subfigures on first column show original images randomly sampled from IDC, Kvasir and HAM10000 validation set. Second to last column are their Grad-CAM visualisations. The visualisations compare all five learning methods, demonstrating visually that TCL is able to generate a localised target lesion location.