CASCADE LEARNING LOCALISES DISCRIMINANT FEATURES IN VISUAL SCENE CLASSIFICATION

Junwen Wang[*]
School of Electronics & Computer Science
University of Southampton
Southampton, UK
SO17 3AT

Katayoun Farrahi
School of Electronics & Computer Science
University of Southampton
Southampton, UK
SO17 3AT

ABSTRACT

Lack of interpretability of deep convolutional neural networks (DCNN) is a well-known problem particularly in the medical domain as clinicians want trustworthy automated decisions. One way to improve trust is to demonstrate the localisation of feature representations with respect to expert-labeled regions of interest. In this work, we investigate the localisation of features learned via two varied learning paradigms and demonstrate the superiority of one learning approach with respect to localisation. Our analysis on medical and natural datasets shows that the traditional end-to-end (E2E) learning strategy has a limited ability to localise discriminative features across multiple network layers. We show that a layer-wise learning strategy, namely cascade learning (CL), results in more localised features. Considering localisation accuracy, we not only show that CL outperforms E2E but that it is a promising method of predicting regions. On the YOLO object detection framework, our best result shows that CL outperforms the E2E scheme by 2% in mAP.

1 Introduction

Deep Learning (DL) advances in computer vision Krizhevsky et al. [2012], Minaee et al. [2022], Redmon and Farhadi [2018] have been successfully applied to specialist domains such as medical imaging Esteva et al. [2021], improving performance in pathology detection from chest radiograph Irvin et al. [2019], Arias-Londono et al. [2020], finding malignant lesions from skin scans Liu et al. [2020] and predicting patient survival from whole slide images Srinidhi et al. [2021]. The success of DL in medical imaging motivates further investigation into feature understanding as these architectures suffer from their black-box nature, raising valid concerns by medical practitioners.

Interpretability is generally the ability for a human to understand the reasons (i.e. features) behind the decision made by the system. Simple machine learning models, such as logistic regression or decision trees, are more easily interpretable though do not perform nearly as well as DCNNs with millions of parameters. Feature visualisation Reyes et al. [2020] is the current state of the art approach for DCNN interpretation. Feature visualisation techniques generate localisation maps, highlighting the pixels and regions in the input image used in making the prediction Saporta et al. [2022], Simonyan et al. [2014], Selvaraju et al. [2020].

Cascade Learning (**CL**) Marquez et al. [2018], which builds on the idea of the cascade correlation algorithm Fahlman and Lebiere [1990], is an alternative way of training a DCNN. This learning paradigm 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 Du et al. [2019], Wang et al. [2022] demonstrate the superior performance of transferring CL features to downstream classification tasks. In this paper, we investigate the difference in feature representations considering localisation as a key metric for traditional E2E learning versus CL. We observe that CL does result in more localised features, considering several metrics and visualisation approaches, and these features appear to be more localised at every layer of the DCNN.
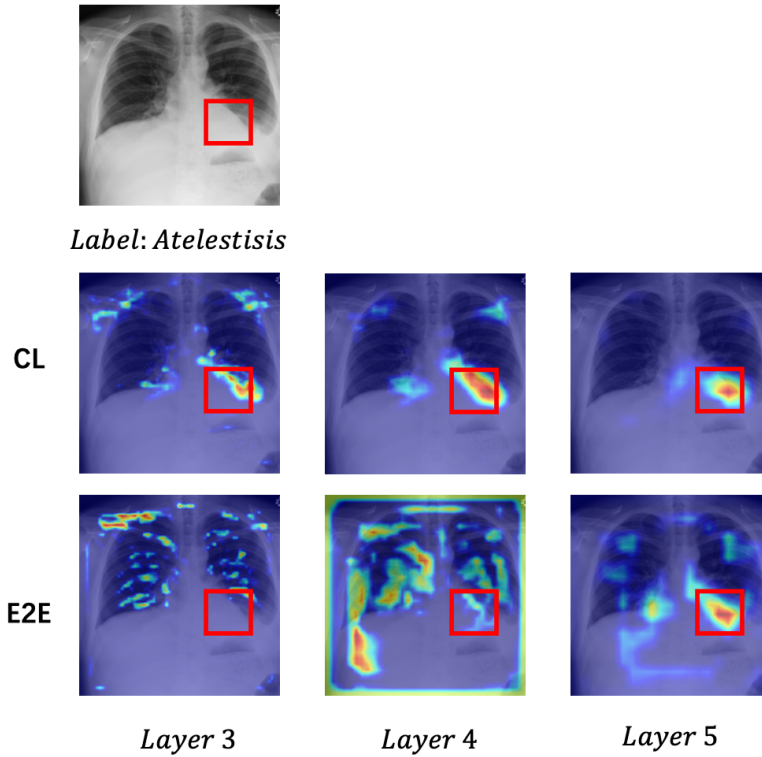
---

[*]junwenwang86@gmail.com

Figure 1: Grad-CAM saliency map visualisation at different layers of the neural network. Results on a (top) cascade-trained network versus (bottom) E2E training. By comparing the features to the red rectangle denoting the bounding box, CL achieves better localisation.

We then take these findings one step further and consider whether the improved feature localisation results in superior object detection. Object detection frameworks train an effective bounding box regressor to classify and localise the object in an image or video Redmon et al. [2016], Redmon and Farhadi [2018]. In this work, we consider the association between visually localised features and the bounding box prediction. We seek to answer: *does the superior localisation ability of CL further improve the ability of the model to predict the bounding box region of interest?* We find that CL is promising and improves bounding box region of interest predictions in comparison to the widely adopted E2E training scheme.

The main contributions of this paper are as follows:

- Our analysis via various feature visualisation techniques shows that traditional E2E training has a limited ability to localise discriminative features across the intermediary layers of a DCNN.

- We demonstrate that using a layer-wise learning strategy, namely cascade learning, leads to an improvement in feature localisation.

- Quantifying the degree of overlap between the binarized mask and the bounding box, for the Chest X-ray dataset, 86% images have more localised features, with CL showing a consistent improvement across every network layer.

- We find the superior localisation ability leads to further improvement in predicting bounding box regions of interest. Our bounding box prediction via CL trained backbone leads to 2% improvement in mAP in object detection tasks.

- We demonstrate that CL learns different features, with coarser features in early layers and finer features in later layers whereas end-to-end learned features have more evenly distributed granulometry across layers.

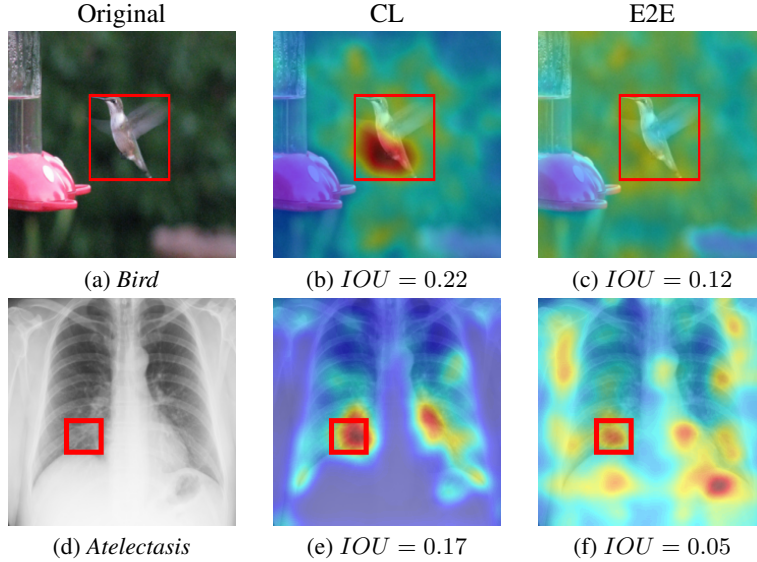|		
(a) *Bird*	(b) $IOU = 0.22$	(c) $IOU = 0.12$
(d) *Atelectasis*	(e) $IOU = 0.17$	(f) $IOU = 0.05$

Figure 2: Saliency map generated via CL in comparison to the same network which is E2E trained. Left column: Original image and its corresponding label; Middle: **CL**; Right: **E2E**. The heatmap was generated after post-processing using a Gaussian filter.

2 Methodology

In this section, we introduce our proposed methodology. Firstly, we describe our technical contribution. Secondly, we describe different techniques in feature visualisation. Thirdly, we briefly describe the YOLO framework Redmon et al. [2016] and how it performs bounding box prediction in object detection tasks. Lastly, we introduce our quantification metric and datasets used in our experiments.

2.1 Deep Cascade Learning in Feature Localisation

One of the important technical contributions of our work is to train the deep neural network from scratch via CL, then perform feature visualisation at different layers and investigate the differences in the feature representations with respect to the labelled bounding box of interest. We perform an identical experimental setup for E2E-trained models. This is partially done by retraining classifiers tapped after every convolutional layer. Our experimental result suggests that E2E-trained models are not localised to the bounding box of interest. Section 2.2 details the feature visualisation methods we adopt in our experiments. Despite the methodology being straightforward, we make an important observation that CL produces high-quality visual explanations compared to identical architectures trained via E2E learning. Our localisation experiment quantitatively demonstrates that feature saliency generated by CL highly overlaps with the region of interest annotated by domain experts. Furthermore, we propose to use CL in DCNN training as an effective bounding box regressor. Our experimental result suggests that DCNN backbone trained via CL improves performance in object detection. Section 2.3 includes the methodology details of the bounding box prediction method.

2.2 Feature Visualisation

Sometimes it is not sufficient to report and be satisfied with strong performance measures on general datasets when delivering care for patients Esteva et al. [2021]. It requires a deep understanding of which cases the model has made a good performance and which circumstances it fails. Feature visualisation provides a visual explanation by plotting salient images showing the most contributing pixel location Simonyan et al. [2014], Selvaraju et al. [2020], or selecting image patches that are potentially interpretable by a model trained via perturbed images Ribeiro et al. [2016].

2.2.1 Saliency Map

Saliency map Simonyan et al. [2014] measures sensitivity for individual pixels, given an input image I on the final prediction. This is achieved by taking the gradient of the class score (S_c) with respect to the input image itself:

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

The result will give us a contribution 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.

2.2.2 Grad-CAM

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 kth feature map and class c. Finally, $L_{\mathrm{Grad-CAM}}$ is computed using 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:

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

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

2.2.3 LIME

Local Interpretable Model-agnostic Explanations (LIME) Ribeiro et al. [2016] generates an occluded version of the image as a visual explanation. This is achieved by randomly perturbing the image patch (allocated by super-pixel) and training simple classifiers (e.g. ridge regression) using prediction score from the model to be explained Ribeiro et al. [2016]. By performing the feature selection on a simple classier, super-pixels that contribute largely to final predictions are found. Figure 3 shows an illustration of the LIME framework Ribeiro et al. [2016].
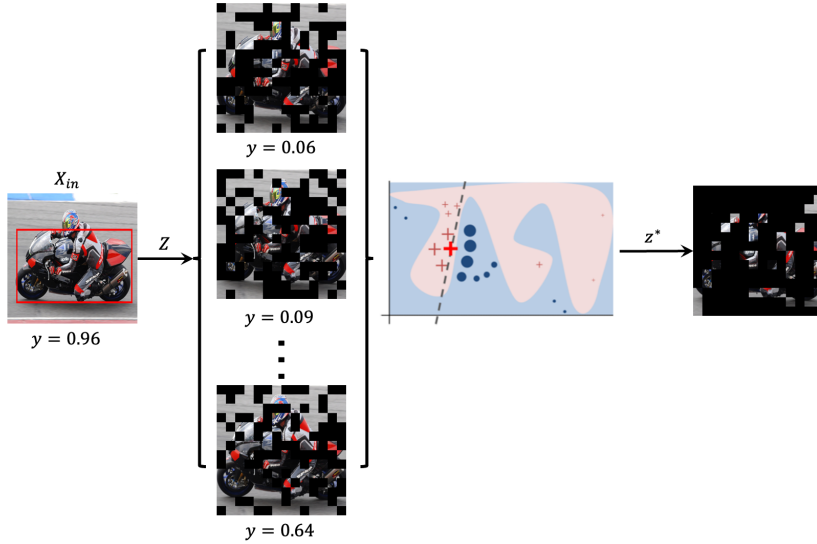

Figure 3: Illustration of the LIME framework. X_{in} is input image; y is confidence score output by model. They introduce a binarized "intermediate representation" z to represent the existence of certain image patches.

2.3 Bounding Box Prediction Via YOLO

You Only Look Once (YOLO) Redmon et al. [2016], Redmon and Farhadi [2018] framework aims to predict all bounding boxes by inputting images once. It splits a given image into an $S \times S$ grid, and predicts the bounding box

coordinate depending on which grid the center of the bounding box is located. It optimizes the following bound-box regression loss:

$$L_{BBox} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$
$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \tag{4}$$

where x, y, w, h represent the two-dimensional object's center coordinate, width and height, respectively. Final loss is calculated by iterating through all grids S and object bounding boxes B. To predict coordinate information, it is normal to modify the output layer or add extra detection head Redmon et al. [2016], Redmon and Farhadi [2018]. However, the network backbone remains the same. In order to implement CL for the object detection task, we consider a two-step approach. First, pre-training the backbone network via CL on image classification tasks. Second, perform bounding-box regression via Equation 4. In our experiments, we consider two backbone network structures, which are a simple 6 layer DCNN model and a 53 layer DarkNet Redmon et al. [2016].

2.4 Metrics

To quantify the localisation ability, we use the *Intercept Over Union* (IOU) metric:

$$IOU = \frac{\text{area} \left(B_p \cap B_{gt} \right)}{\text{area} \left(B_p \cup B_{gt} \right)} \tag{5}$$

where B_p denotes the binarized saliency map. For the thresholding process, we use a fixed percentile instead of a constant value, ensuring a fair comparison. This results in binarized saliency maps that all have the same degree of pixel covering but are different in distribution. B_{gt} denotes the binarized ground truth bounding box, where regions inside the box are *True*. To quantify the model's overall localisation ability, we define *Localisation Accuracy* by measuring the fraction of instances that satisfy $IOU > 0.2$. Note that the LIME framework explains the decision at the patch level. However, we are merely interested in part of the patch that overlaps with the bounding box. Therefore, we measure mainly the degree of overlap by counting the number of pixels that are inside the bounding box.

For the object detection task, we evaluate our model performance using *mean Average Precision (mAP)* Lin et al. [2014] and *mean Intersection over Union (mIOU)*. We are using both single and multiple IOU thresholds to measure mAP. For a single IOU threshold, we select $IOU = 0.5$ and 0.75. For multiple IOU thresholds, we use the mean of 10 IOU thresholds, from 0.5 to 0.95 with step size 0.05.

2.5 Datasets

We show our method has improved localisation ability in both natural image and medical domains. Specifically, for the natural image domain, we choose Pascal VOC Everingham et al. [2010] which includes $11,530$ natural images in 20 classes and $27,450$ object ROI since multiple objects exist. For chest X-ray images we use the ChestX-ray8 Wang et al. [2017] dataset, where 987 chest X-ray images are provided with board certified medic annotations of the correct location of the anomaly.

3 Results

Next, we present experimental results on the two varied datasets of natural and medical images.

3.1 Feature Visualisation

In Figure 1, we visualise the dominant features learned by the network across various layers using the Grad-CAM Selvaraju et al. [2020] saliency map. We observe a large gap between CL (top row) and E2E (bottom row) features on the chest X-ray data. CL features are often more visually localised with respect to the bounding box. Similar phenomena are observed by only visualising the gradient signals across various layers as illustrated in Figure 2. These results suggest that the gradient signal plays an important role in generating a qualitative visualisation. Next, we quantify this effect by considering the IOU and localisation accuracy over both datasets.

3.2 Feature Localisation via Grad-CAM and Saliency Map

Figure 4 shows the scatter plots of IOU computed over (a) 2000 Pascal images and (b) 987 chest X-ray images. Each data point corresponds to an image, with the IOU of the network trained with E2E presented on the x-axis, and CL on the y-axis. The majority of the images have more localised features (higher IOU) with CL as opposed to E2E, with 74% on the Pascal dataset and 86% with the Chest x-ray dataset. The localisation accuracy is further plotted over the layers of the network in Figure 4(c) and (d) demonstrating the superiority in feature localisation for networks trained via CL.

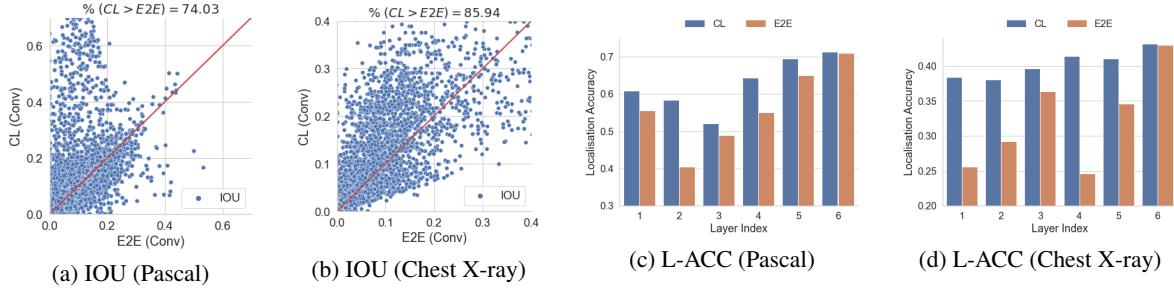(a) IOU (Pascal) (b) IOU (Chest X-ray) (c) L-ACC (Pascal) (d) L-ACC (Chest X-ray)

Figure 4: a-b): Scattering plot of IOU between the manual annotation and saliency maps. The experiment was conducted on both natural images (Pascal) and medical datasets (Chest X-ray). c-d): IOU between the *Grad-CAM* and bounding boxes, over varied learning method layers.

3.3 Feature Localisation via LIME Framework

In this section, we evaluate CL localisation performance using LIME Ribeiro et al. [2016]. LIME requires learning multiple simple learners which creates extra complexity. But it does not require gradient information, which differs from a gradient-based method such as Grad-CAM Selvaraju et al. [2020] and saliency map Simonyan et al. [2014]. In Figure 5, we show that CL produces meaningful features by measuring the degree of overlap between LIME output images (occluded area are treated as 0) and the bounding box. Figure 5 shows localisation performance comparing CL and E2E learning methods using the LIME test. CL learned features consistently outperform the E2E features in all layers, with the largest improvement found at the second layer.
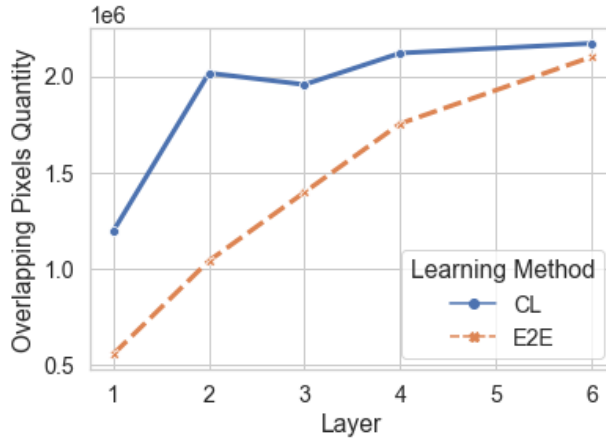

Figure 5: Localisation performance comparing CL and E2E learning method using LIME framework.

3.4 CL Improves Region Proposal

We show that training a backbone network via CL improves region proposal in bounding box prediction. We adopt a network trained via CL as an effective bounding box regressor. We keep feature layers frozen and retrain one added convolutional layer to optimize the bounding box regression loss. The loss was first introduced in YOLOv1 framework Redmon et al. [2016]. The layerwise comparison results are shown in Figure 6. We found CL achieves the best overall quality of region proposal with the largest difference at layer 4 compared to an identical network trained via

E2E. Notice that we are not directly competing with the state-of-the-art model, but we claim using the CL training scheme improves overall performance in object detection tasks against the widely adopted E2E scheme.
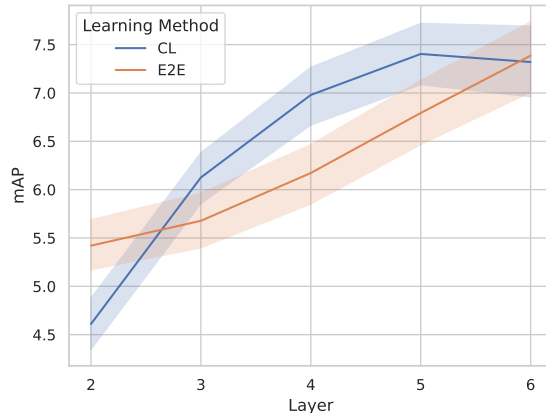

Figure 6: CL achieves better performance with the largest difference in mAP at layer 4. For each layer, we re-train CL in three different random seeds. The shaded area denotes the standard deviation.

We further investigate whether using a deep network backbone trained via CL could improve region proposal. Table 1 shows YOLOv3 performance on the Pascal dataset. We adopt CL to train the DarkNet-53 Redmon and Farhadi [2018] network backbone from scratch and compare against the E2E baseline training method. To implement the CL algorithm on the DarkNet-53 architecture, we split the whole structure into multiple sub-modules. Each sub-module consists of at least one complete residual connection block. When the network architecture is fixed, the size of the sub-module is determined by the total number of splits. In the YOLOv3 framework, the output is taken from three locations among intermediate features and passed to the feature pyramid network (FPN) to improve detection for different object sizes. In our experiment, we split the network into three sub-modules and denote it as CL_3. This result, along with other splitting strategies to train with CL, are reported in Table 1. We found using pre-trained features from the middle layer of the network yields the largest difference between CL and E2E. The best performance using CL feature up to middle layer improves 2% in $mAP_{.5}$ metric compared to reusing E2E feature at same layer. When increasing the number of splits, the performance starts to decrease. This is possibly caused by overfitting since the network's learning ability is limited due to sub-module size shrinks by having larger quantities of splits.

	$mAP_{.5:.95:.05}$	$mAP_{.5}$	$mAP_{.75}$
CL_3	34.64±0.31	67.29±0.28	31.77±0.4
CL_7	33.87±0.1	66.56±0.15	30.27±0.3
CL_{23}	33.34±0.28	65.68±0.23	29.63±0.62
$E2E$	33.41±0.18	65.3±0.24	30.07±0.43

Table 1: Comparing CL and E2E training method to train network backbone in YOLOv3 framework Redmon and Farhadi [2018]. All CL and E2E are using DarkNet-53 Redmon and Farhadi [2018] architecture. The lower subscript denotes the total number of splits in CL training.

3.5 Quantifying Coarse-to-Fine Features Representation

Granulometry analysis Dougherty et al. [1989] on the generated saliency maps quantitatively demonstrates the coarse-to-fine feature representation. The higher granulometry represents the feature activation (indicated as the irregular red patch in Figure 1) are coarser, finer detail is learned if granulometry has a low value. Figure 7 quantitatively analyze using granulometry to measure CL and E2E feature representation. We conclude that CL is learning coarser feature representation at early layers and finer at later layers. On the contrary, E2E has more evenly distributed granulometry across the layers. These results strengthen the argument for CL learning optimal feature representation as we demonstrate that early layers in the network are learning coarser features while later layers are learning more fine-grained features.
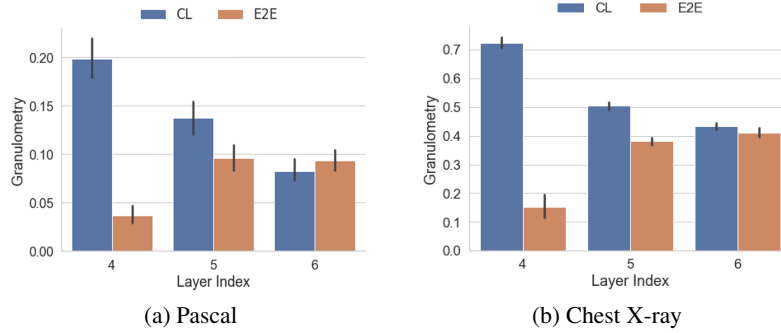
(a) Pascal

(b) Chest X-ray

Figure 7: Granulometry measure comparing CL and E2E learning methods on different layers (layer $1 - 3$ as the inconsistency observed in early layers). a) Pascal; b) chest X-ray

3.6 Visualisation for Small Object Bounding Box

In Figure 8, we visualise instances with a relatively small bounding box. By visualising the instance and its corresponding binary mask, we observed that CL is able to generate a localised heatmap for the small object of interest with a high-quality salient image. On the other hand, E2E tend to generate salient images that are activated in a large region, result in an ambiguous localisation.
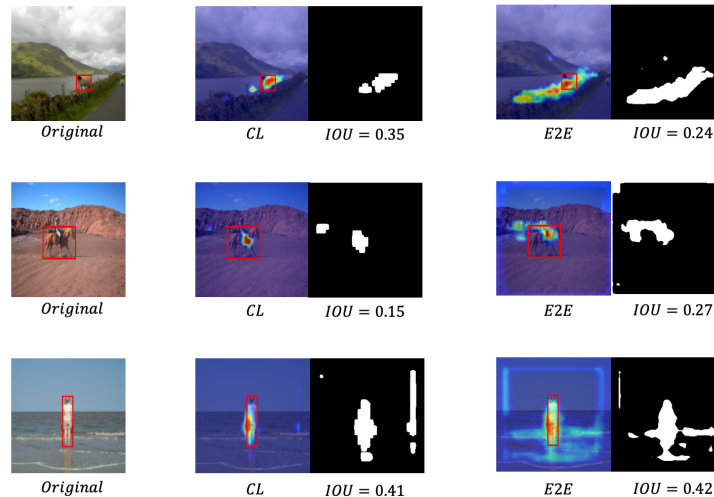

Figure 8: Visualisation of data instance via Grad-CAM generated from two methods. left: Original image and its bounding box; middle: CL ; right: E2E. The IOU value associated with each binarized saliency map is shown on the right side.

4 Conclusion

In this work we investigate the localisation of features across learning paradigms. Our systematic evaluation across various feature visualisation methods and datasets show that E2E training, which has been widely considered by the machine learning community, is limited to localising discriminative features across multiple network layers. We found network trained via CL is more localised to the region of interest annotated by domain experts. We show that CL's superior localisation ability leads to an improvement in object detection tasks.

References

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, pages 1097–1105. Curran Associates Inc., 2012.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (7):3523–3542, 2022.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, et al. Deep Learning-enabled Medical Computer Vision. *npj Digital Medicine*, 4(1):5, 2021.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Ciurea-Ilcus, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. 2019.

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*, 2020. doi:10.1109/ACCESS.2020.3044858.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, 2020. ISSN 1546-170X.

Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. ISSN 1361-8415.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, et al. On The Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, 2020.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, et al. Benchmarking saliency methods for chest X-ray interpretation. *medRxiv*, 2022. doi:10.1101/2021.02.28.21252634.

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

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128: 336–359, 2020.

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.

Scott E. Fahlman and Christian Lebiere. The Cascade-Correlation Learning Architecture. In *Advances in Neural Information Processing Systems (NIPS)*, page 524–532, 1990. ISBN 1558601007.

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.

Junwen Wang, Xin Du, Katayoun Farrahi, and Mahesan Niranjan. Deep Cascade Learning for Optimal Medical Image Feature Representation. In *Machine Learning for Healthcare (MLHC)*, 2022.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. pages 779–788, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144. Association for Computing Machinery, 2016.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge". *International Journal of Computer Vision*, 88(2):303–338, June 2010.

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. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3462–3471, 2017.

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.

5 Appendix

5.1 More Results

Figure 9 shows scatter plots of IOU and localisation accuracy. Quantifying the localisation ability of CL via *Saliency Map* Simonyan et al. [2014]. Align with the result in Figure 4, CL learning scheme consistently improves localisation over E2E learning scheme.

(a) IOU (Pascal)   (b) IOU (Chest X-ray)   (c) L-ACC (Pascal)   (d) L-ACC (Chest X-ray)
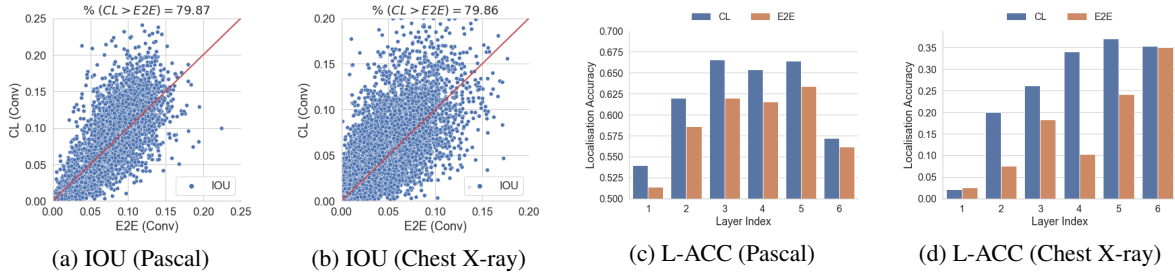
Figure 9: a-b): Scattering plot of IOU between the manual annotation and saliency maps. The experiment was conducted on both natural image (pascal) and medical dataset (chest X-ray) c-d): IOU between the *saliency maps* output and bounding boxes, over different layers.

Figure 10 provides qualitative analysis by visualising bounding-box prediction for some randomly selected images. We notice CL is able to predict a precise bounding box location. On the other hand, E2E fails to generate the bounding box (e.g. second row, image of jar) or generate imprecise location (e.g. first row, image of two cats).

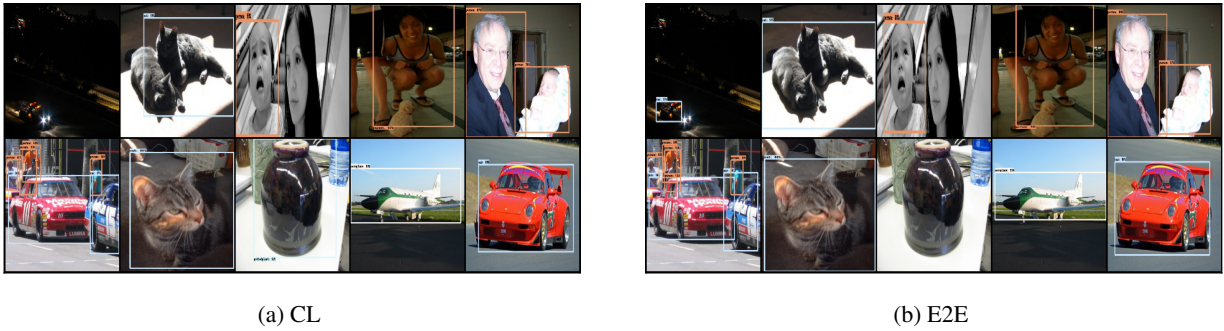(a) CL   (b) E2E

Figure 10: Qualitative Results. YOLO prediction on random test sample from Pascal dataset. Comparing CL and E2E training scheme to train network backbone.