Jointly Boosting Saliency Prediction and Disease Classification on Chest X-ray Images with Multi-task UNet

Hongzhi Zhu[1(✉)], Robert Rohling[1,2,3], and Septimiu Salcudean[1,2]

[1] School of Biomedical Engineering, University of British Columbia, Vancouver, Canada
{hzhu,rohling,tims}@ece.ubc.ca
[2] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada
[3] Department of Mechanical Engineering, University of British Columbia, Vancouver, Canada

Abstract. Human visual attention has recently shown its distinct capability in boosting machine learning models. However, studies that aim to facilitate medical tasks with human visual attention are still scarce. To support the use of visual attention, this paper describes a novel deep learning model for visual saliency prediction on chest X-ray (CXR) images. To cope with data deficiency, we exploit the multi-task learning method and tackle disease classification on CXR simultaneously. For a more robust training process, we propose a further optimized multi-task learning scheme to better handle model overfitting. Experiments show our proposed deep learning model with our new learning scheme can outperform existing methods dedicated either for saliency prediction or image classification. The code used in this paper is available at [webpage, concealed for double-blind review].

Keywords: Saliency prediction · Disease classification · X-ray imaging · Deep learning · Multi-task learning

1 Introduction

Recent work in machine learning and computer vision have demonstrated advantages of integrating human attention with artificial neural network models, as studies show that many machine vision tasks, i.e., image segmentation, image captioning, object recognition, etc., can benefit from adding human visual attention [36].

Visual attention is the ability inherited in biological visual systems to selectively recognize regions or features on scenes relevant to a specific task [3], where "bottom-up" attention (also called exogenous attention) focuses on physical properties in the visual input that are salient and distinguishable, and "top-down" attention (also called endogenous attention) generally refers to mental

G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 594–608, 2022.
https://doi.org/10.1007/978-3-031-12053-4_44

strategies adopted by the visual systems to accomplish the intended visual tasks [44]. Early research on saliency prediction aims to understand attentions triggered by visual features and patterns, and thus "bottom-up" attention is the research focus [3]. More recent attempts, empowered by interdisciplinary efforts, start to study both "bottom-up" and "top-down" attentions, and therefore the terms, saliency prediction and visual attention prediction, are used interchangeably [53]. In this paper, we use the term saliency prediction as the prediction of human visual attentions allocations when viewing 2D images, containing both "bottom-up" and "top-down" attentions. 2D heatmap is usually used to represent human visual attention distribution. Note that saliency prediction studied in this paper is different from neural network's saliency/attention which can be visualized through class activation mapping (CAM) by [63] and other methods [15,48,51]. With the establishment of several benchmark datasets, data driven approaches demonstrated major advancements in saliency prediction (review in [2] and [60]). However, saliency prediction for natural scenes is the primary focus, and more needs to be done in the medical domain. Hence, we intend to study the saliency prediction for examining chest X-ray (CXR) images, one of the most common radiology tasks worldwide.

CXR imaging is commonly used for the diagnosis of cardio and/or respiratory abnormalities; it is capable of identifying multiple conditions through a single shot, i.e., COVID-19, pneumonia, heart enlargement, etc. [6]. There exists multiple public CXR datasets [20,61]. However, the creation of large comprehensive medical datasets is labour intensive, and requires significant medical resources which are usually scarce [9]. Consequently, medical datasets are rarely as abundant as those for non-medical fields. Thus, machine learning approaches applied on medical datasets need to address the problem of data scarcity. In this paper, we exploit the multi-task learning for a solution.

Multi-task learning is known for its inductive transfer characteristics that can drive strong representation learning and generalization of each component task [8]. Therefore, multi-task learning methods partially alleviates some of the major shortcomings in deep learning, i.e., high demands for data sufficiency and heavy computation loads [11]. However, to apply multi-task learning methods successfully, challenges still exist, which can be the proper selection of component tasks, the architecture of the network, the optimization of the training schemes and many others [11,62]. This paper investigates the proper configuration of a multi-task learning model that can tackle visual saliency prediction and image classification simultaneously.

The main contributions of this paper are: 1) development of a new deep convolutional neural network (DCNN) architecture for CXR image saliency prediction and classification based on UNet [47], and 2) proposal of an optimized multi-task learning scheme that handles overfitting. Our method aims to outperform the state-of-the-art networks dedicated either for saliency prediction or image classification.

2 Background

2.1 Saliency Prediction with Deep Learning

DCNN is the leading machine learning method applied to saliency prediction [22,30,31,43]. Besides, transfer learning with pre-trained networks was observed to boost the performance of saliency prediction [31,41,42]. A majority of DCNN approaches are for natural scene saliency prediction, and so far, only a few studied the saliency prediction for medical images. By [5], the generative adversarial network is used to predict expert sonographer's saliency when performing standard fetal head plane detection on ultrasound (US) images. However, the saliency prediction is used as a secondary task to assist the primary detection task, and thus, the saliency prediction performance failed to outperform benchmark prediction methods in several key metrics. Similarly, by [25], as a proof-of-concept study, the gaze data is used as an auxiliary task for CXR image classification, and the performance of saliency prediction is not reported in the study.

2.2 CXR Image Classification with Deep Learning

Public datasets for CXR images enabled data driven approaches for automatic image analysis and diagnosis [33,50]. Advancements in standardized image classification networks, i.e., ResNet [18], DenseNet [19], and EfficientNet [55], facilitate CXR image classification. Yet, CXR image classification remains challenging, as CXR images are noisy, and may contain subtle features that are difficult to recognize even by experts [6,28].

3 Multi-task Learning Method

As stated in Sect. 1, component task selection, network architecture design, and training scheme are key factors for multi-task learning. We select the classification task together with the saliency prediction based on the fact that attention patterns are task specific [26]. Radiologists are likely to exhibit distinguishable visual behaviors when different patient conditions are shown on CXR images [38]. This section introduces our multi-task UNet (MT-UNet) architecture, and derives a better multi-task training scheme for saliency prediction and image classification.

3.1 Multi-task UNet

Figure 1 shows the architecture of the proposed MT-UNet. The network takes CXR images, $\boldsymbol{x} \in \mathbf{R}^{1\times H\times W}$, where H and W are image dimensions, as input, and produces two outputs, predicted saliency $\boldsymbol{y}_s \in \mathbf{R}^{1\times H\times W}$, and predicted classification $\boldsymbol{y}_c \in \mathbf{R}^{C}$, where C is the number of classes. As the ground truth for $\boldsymbol{y}_s$ is human visual attention distribution, represented as a 2D matrix whose elements are non-negative and sum to 1, $\boldsymbol{y}_s$ is normalized by Softmax before

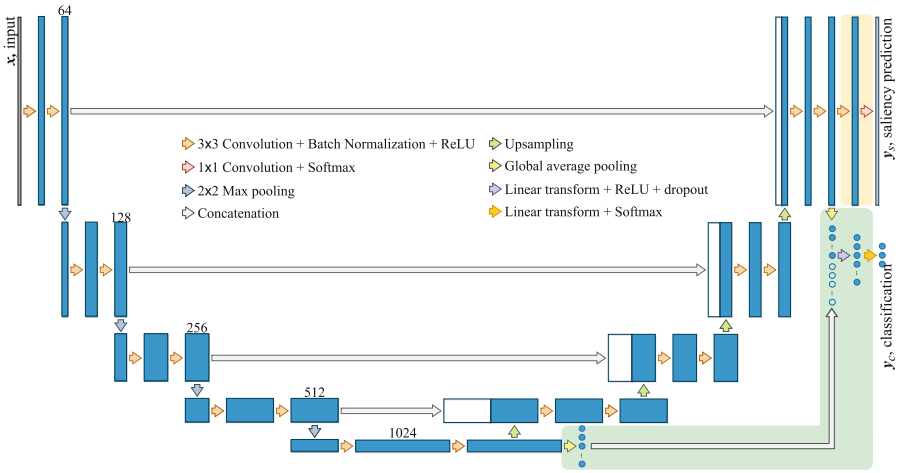


Fig. 1. MT-UNet architecture. The solid blocks represent 3D tensors, $\mathbf{R}^{F\times H\times W}$, where F, H, and W denote feature (channel), height and width dimensions, respectively. The solid circles represent 1D tensors. Arrows denote operations to the tensors. Numbers above some of the solid blocks stand for the number features in tensors.

output from MT-UNet. Softmax is also applied to $\boldsymbol{y}_c$ before output so that the classification outcome can be interpreted as class probability. For the simplicity of notation, batch dimensions are neglected.

The proposed MT-UNet is derived from standard UNet architecture [47]. As a well-known image-to-image deep learning model, the UNet structure has been adopted for various tasks. For example, the UNet is appended with additional structures for visual scene understanding [21], the features from the bottleneck (middle of the UNet) are extracted for image classification tasks [25], and by combining UNet with Pyramid Net [35], features at different depth are aggregated for enhanced segmentation [40]. What's more, the encoder-decoder structure of UNet is utilized for multi-task learning, where the encoder structure is used to learn representative features, along with designated decoder structures or classification heads for image reconstruction, segmentation, and/or classification [1,64]. In our design, we apply classification heads (shaded in light green in Fig. 1), which are added not only to the bottleneck but also the ending part of the UNet architecture. This additional classification specific structure aggregates middle and higher-level features for classification, exploiting features learnt at different depths. The attention heads perform global average pooling operations to the 4D tensors, followed by concatenation, and two linear transforms (dense layers) with dropout (rate = 25%) in the middle to produce classification outcomes. The MT-UNet belongs to the hard parameter sharing structure in multi-task learning, where different tasks share the same trainable parameters before branched out to each tasks' specific parameters [58]. Having more trainable parameters in task specific structures may improve the performance for that

task at the cost of introducing additional parameters and increasing computational load [11,58]. In our design, we wish to avoid heavy structures with lots of task specific parameters, and therefore, task specific structures are minimized. In Fig. 1, we use yellow and green shades to denote network structures dedicated for saliency prediction and classification, respectively.

3.2 Multi-task Training Scheme

Balancing the losses between tasks in a multi-task training process has a direct impact on the training outcome [58]. There exist multi-task training schemes [10, 16,27,49], and among which, we adopt the uncertainty based balancing scheme [27] with the modification used in [34,65]. Hence, the loss function is:

$$\mathcal{L} = \frac{1}{\sigma_s^2} L_s + \frac{1}{\sigma_c^2} L_c + \ln(\sigma_s + 1) + \ln(\sigma_c + 1) \tag{1}$$

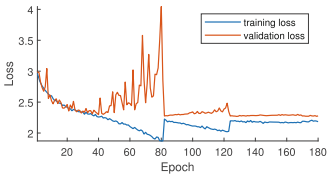
where L_s and L_c are loss values for $\boldsymbol{y}_s$ and $\boldsymbol{y}_c$, respectively; $\sigma_s > 0$ and $\sigma_c > 0$ are trainable scalars estimating the uncertainty of L_s and L_c, respectively; σ_s and σ_c are initialized to 1; $\ln(\sigma_s + 1)$ and $\ln(\sigma_c + 1)$ are regularizing terms to avoid arbitrary decrease of σ_s and σ_c. With Eq. 1, we know that σ values can dynamically weigh losses of different amplitudes during training, and loss with low uncertainty (small σ value) is prioritized in the training process. $\mathcal{L} > 0$. Given $\boldsymbol{y}_s$ and $\boldsymbol{y}_c$ with their ground truth $\bar{\boldsymbol{y}}_s$ and $\bar{\boldsymbol{y}}_c$, respectively, the loss functions are:

$$L_s = H(\bar{\boldsymbol{y}}_s, \boldsymbol{y}_s) - H(\bar{\boldsymbol{y}}_s), \tag{2}$$

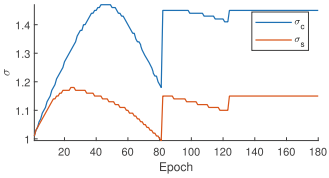
$$L_c = H(\bar{\boldsymbol{y}}_c, \boldsymbol{y}_c) \tag{3}$$

where $H(Q, R) = -\Sigma_i^n Q_i \ln(R_i)$ stands for cross entropy of two discrete distributions Q and R, both with n elements; $H(Q) = H(Q, Q)$ stands for the entropy, or self cross entropy, of discrete distribution Q. L_s is the Kullback-Leibler divergence (KLD) loss, and L_c is the cross-entropy loss. By observing Eq. 2 and Eq. 3, we know that only the cross entropy terms, $H(\cdot, \cdot)$, generate gradient when updating network parameters, as the term $-H(\bar{\boldsymbol{y}}_s)$ in L_s is a constant and has zero gradient. Therefore, we extend the method in [27], and use $\frac{1}{\sigma^2}$ to scale a KLD loss (L_s) as that for a cross-entropy loss (L_c).

Although the training scheme in Eq. 1 yields many successful applications, overfitting for multi-task networks still can jeopardize the training process, especially for small datasets [59]. Multiple factors can cause overfitting, among which, learning rate, $r > 0$, shows the most significant impact [32]. Also, r generally has significant influences on the training outcome [52], making it one of the most important hyper-parameters for a training process. When training MT-UNet, r is moderated by several factors. The first factor is the use of an optimizer. Many optimizers, i.e., Adam [29] and RMSProp [57], deploy the momentum mechanism or its variants, which can adaptively adjust the effective learning rate, r_e, during training. As a learning rate scheduler is often used for more efficient training, it is the second factor to influence r. The influence of r from a learning rate



(a) Losses



(b) σ values

Fig. 2. Training process visualization with Eq. 1

scheduler can be adaptive, i.e., reduce learning rate on plateau (RLRP), or more arbitrary, i.e., cosine annealing with warm restarts [37]. By observing Eq. 1, we know that an uncertainty estimator σ for a loss L also serves as a learning rate adaptor for L, which is the third factor. More specifically, given a loss value L with learning rate r, the effective learning rate for parameters with a scaled loss value $\frac{L}{\sigma^2}$ is $\frac{r}{\sigma^2}$.

Decreasing r upon overfitting can alleviate its effects [12,52], but Eq. 1 leads to increased learning rate upon overfitting, further worsening the training process. This happens because training loss decreases when overfitting occurs, reducing its variance at the same time. Thus, σ decreases accordingly, which increases the effective learning rate, thus creating a vicious circle of overfitting. This phenomenon can be observed in Fig. 2, where changes of losses and σ values during a training process following Eq. 1 are presented. We can see from Fig. 2(a), at epoch 40, after an initial decrease in both the training and validation losses, the training loss start to decrease acceleratedly while the validation loss start to amplify, which is a vicious circle of overfitting. A RLRP scheduler can halt the vicious circle by resetting the model parameters to a former epoch and reducing r. Yet, even with reduced r, a vicious circle of overfitting can remerge in later epochs. The mathematical proof of the aforementioned vicious circle of overfitting is presented in Appendix A.

To alleviate overfitting, we propose the use of the following equations to replace Eq. 1:

$$\mathcal{L} = \frac{1}{\sigma_s^2} L_s + L_c + \ln(\sigma_s + 1), \tag{4}$$

$$\mathcal{L} = L_s + \frac{1}{\sigma_c^2} L_c + \ln(\sigma_c + 1). \tag{5}$$

The essence of Eqs. 4 and 5 is to fix the uncertainty term for one loss in Eq. 1 to 1, so that the flexibility in changing effective learning rate is reduced. With the uncertainty term fixed for one component loss, Eqs. 4 and 5 demonstrate the ability to alleviate overfitting and stabilize the training process. It is worth noting that Eqs. 4 and 5 cannot be used interchangeably. We need to test both equations to check which can achieve better performances, as depending on the dataset and training process, overfitting can occur of different severity in all component tasks. In this study, the training process with Eq. 5 achieves the best performance. Ablation study of this method is presented in Sect. 5.

4 Dataset and Evaluation Methods

We use the "chest X-ray dataset with eye-tracking and report dictation" [25] shared via PhysioNet [39] in this study. The dataset was derived from the MIMIC-CXR dataset [23,24] with additional gaze tracking and dictation from an expert radiologist. 1083 CXR images are included in the dataset, and accompanying each image, there are tracked gaze data; a diagnostic label (either normal, pneumonia, or enlarged heart); segmentation of lungs, mediastinum, and aortic knob; and radiologist's audio with dictation. The CXR images in the dataset are in resolutions of various sizes, i.e., 3056×2044, and we down sample and/or pad each image to 640×416. A GP3 gaze tracker by Gazepoint (Vancouver, Canada) was used for the collection of gaze data. The tracker has an accuracy of around 1° of visual angle, and has a 60 Hz sampling rate [66].

Several metrics have been used for the evaluation of saliency prediction performances, and they can be classified into location-based metrics and distribution-based metrics [4]. Due to the tracking inaccuracy of the GP3 gaze tracker, location-based metrics are not suited for this study. Therefore, in this paper, we follow suggestions in [4] and use KLD for performance evaluation. We also include histogram similarity (HS), and Pearson's correlation coefficient (PCC) for reference purposes. For the evaluation of classification performances, we use the area under curve (AUC) metrics for multi-class classifications [14,17], and the classification accuracy (ACC) metrics. We also include the AUC metrics for each class: normal, enlarged heart, and pneumonia, denoted as AUC-Y1, AUC-Y2, and AUC-Y3, respectively. In this paper, all metrics values are presented as median statistics followed by standard deviations behind the $\pm$ sign. Metrics with up-pointing arrow $\uparrow$ indicates greater values reflect better performances, and vice versa. Best metrics are emboldened.

5 Experiments and Result

5.1 Benchmark Comparison

In this subsection, we compare the performance of MT-UNet, with benchmark networks for CXR image classification and saliency prediction. Detailed training settings are presented in Appendix B.

For CXR image classification, the benchmark networks are chosen from the top performing networks for CXR image classification examined in [13], which are ResNet50 [18] and Inception-ResNet v2 (abbreviated as IRNetV2 in this paper) [54]. Following [25], we also include a state-of-the-art general purpose classification network: EfficientNetV2-S (abbreviated as EffNetV2-S) [56] for comparison. For completeness, classification using standard UNet with additional classification head (denoted as UNetC) is included. Results are presented in Table 1, and We can see that MT-UNet outperforms the other classification networks.

For CXR image saliency prediction, comparison was conducted with 3 state-of-the-art saliency prediction models, which are SimpleNet [46], MSINet [30] and VGGSSM [7]. Saliency prediction using standard UNet (denoted as UNetS) is

Table 1. Performance comparison between classification models.

Metrics	MT-UNet	UNetC	EffNetv2-S	IRNetv2	ResNet50
ACC ↑	**0.670** ± 0.018	0.593 ± 0.009	0.640 ± 0.037	0.640 ± 0.017	0.613 ± 0.013
AUC ↑	**0.843** ± 0.012	0.780 ± 0.006	0.826 ± 0.015	0.824 ± 0.014	0.816 ± 0.010
AUC-Y1 ↑	**0.864** ± 0.014	0.841 ± 0.007	0.852 ± 0.013	0.862 ± 0.016	0.845 ± 0.015
AUC-Y2 ↑	**0.912** ± 0.008	0.840 ± 0.003	0.901 ± 0.015	0.897 ± 0.011	0.896 ± 0.015
AUC-Y3 ↑	**0.711** ± 0.027	0.597 ± 0.018	0.653 ± 0.017	0.633 ± 0.036	0.622 ± 0.022

Table 2. Performance comparison between saliency prediction models.

Metrics	MT-UNet	UNetS	SimpleNet	MSINet	VGGSSM
KLD ↓	**0.726** ± 0.004	0.750 ± 0.002	0.758 ± 0.009	0.748 ± 0.003	0.743 ± 0.007
PCC ↑	**0.569** ± 0.004	0.552 ± 0.002	0.545 ± 0.008	0.557 ± 0.002	0.561 ± 0.005
HS ↑	**0.548** ± 0.001	0.540 ± 0.001	0.541 ± 0.002	0.545 ± 0.001	0.545 ± 0.003

also included for reference. Table 2 shows the result, where MT-UNet outperforms the rest. Visual comparisons for saliency prediction results are presented through Table 4 in Appendix C.

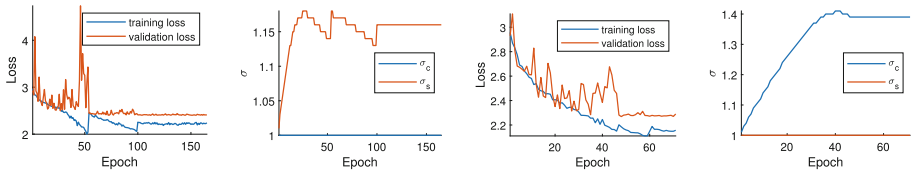
5.2 Ablation Study

To validate the modified multi-task learning scheme, ablation study is performed. The multi-task learning schemes following Eqs. 1, 4, and 5 are compared, and they are denoted as MTLS1, MTLS2, and MTLS3, respectively. Please note that the best-performing MTLS3 is used for benchmark comparison in Sect. 5.1. Figure 3 shows the training process for MTLS2 and MTLS3. With Figs. 2 and 3, we can see that overfitting occurs both for MTLS1 and MTLS2, but the overfitting is reduced in MTLS3. The training processes shown in Figs. 2 and 3 are with optimized hyper-parameters. The resulting performances are compared in Table 3. We can see that MTLS3 outperforms the rest learning schemes both in classification and in saliency prediction.

To validate the effects of using classification head that aggregates features from different depths, we create ablated versions of MT-UNet that use features from either the bottleneck or the top layer of the MT-UNet for classification, denoted as MT-UNetB and MT-UNetT, respectively. Results are presented in Table 3. We can see that MT-UNet generally performs better than MT-UNetT and MT-UNetB.

Table 3. Ablation study performance comparison.

Metrics	MTLS1	MTLS2	MTLS3	MT-UNetB	MT-UNetT
KLD ↓	0.730 ± 0.007	0.738 ± 0.006	**0.726** ± 0.004	0.730 ± 0.003	0.734 ± 0.007
CC ↑	0.566 ± 0.005	0.563 ± 0.005	**0.569** ± 0.004	0.568 ± 0.003	0.561 ± 0.007
HS ↑	0.547 ± 0.002	0.545 ± 0.002	**0.548** ± 0.001	**0.548** ± 0.001	0.544 ± 0.003
ACC ↑	0.649 ± 0.041	0.638 ± 0.019	**0.670** ± 0.018	0.653 ± 0.013	0.649 ± 0.011
AUC ↑	0.832 ± 0.019	0.832 ± 0.010	0.843 ± 0.012	0.836 ± 0.009	**0.847** ± 0.008
AUC-Y1 ↑	0.859 ± 0.014	0.861 ± 0.015	0.864 ± 0.014	0.859 ± 0.007	**0.883** ± 0.005
AUC-Y2 ↑	0.906 ± 0.016	0.913 ± 0.005	**0.912** ± 0.008	0.907 ± 0.011	0.910 ± 0.006
AUC-Y3 ↑	0.682 ± 0.035	0.672 ± 0.010	**0.711** ± 0.027	0.694 ± 0.023	0.695 ± 0.025



(a) MTLS2 losses (b) MTLS2 σ (c) MTLS3 losses (d) MTLS3 σ

Fig. 3. Multi-task learning schemes comparison

6 Discussion

In this paper, we build the MT-UNet model and propose a further optimized multi-tasking learning scheme for saliency prediction and disease classification with CXR images. While a multi-task learning model has the potential of enhancing the performances for all component tasks, a proper training scheme is one of the key factors to fully unveil its potentiality. As shown in Table 3, MT-UNet with the standard multi-task learning scheme may barely outperform existing models for saliency prediction or image classification.

Several future work could be done to improve this study. The first would be the expansion of the gaze tracking dataset for medical images. So far, only 1083 CXR images are publicly available with radiologist's gaze behavior, limiting extensive studies of gaze-tracking assisted machine learning methods in the medical field. Also, more dedicated studies on multi-task learning methods, especially for small datasets, can be helpful for medical machine learning tasks. Overfitting and data deficiency are the lingering challenges encountered by many studies. A better multi-task learning method may handle these challenges more easily.

A Mathematical Derivation of Vicious Circle for Overfitting

Let $L \geq 0$ be the loss for a task, $\mathcal{T}$, and $\sigma > 0$ be the variance estimator for L used in Eq. 1. Therefore, the loss for $\mathcal{T}$ following Eq. 1 can be expressed as:

$$\mathcal{L} = \frac{L}{\sigma^2} + \ln(\sigma + 1). \tag{6}$$

The partial derivative of $\mathcal{L}$ with respect to σ is:

$$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{2L}{\sigma^3} + \frac{1}{\sigma + 1}. \tag{7}$$

During a gradient based optimization process, to minimize $\mathcal{L}$, σ converges to the equilibrium value (σ remains unchanged after gradient descend) which is achieved when $\frac{\partial \mathcal{L}}{\partial \sigma} = 0$. Therefore, the following equation holds when σ is at its equilibrium value, denoted as $\tilde{\sigma}$:

$$L = \frac{\tilde{\sigma}^3}{2\tilde{\sigma} + 2} \tag{8}$$

which is calculated by letting $\frac{\partial \mathcal{L}}{\partial \sigma} = 0$. Let $f(\tilde{\sigma}) = L$, $\tilde{\sigma} > 0$, we can calculate that:

$$\frac{df(\tilde{\sigma})}{d\tilde{\sigma}} = \frac{\tilde{\sigma}^2(2\tilde{\sigma} + 3)}{2(\tilde{\sigma} + 1)^2} > 0, \quad \forall \tilde{\sigma} > 0. \tag{9}$$

Therefore, we know that $f(\tilde{\sigma})$ is strictly monotonically increasing with respect to $\tilde{\sigma}$, and hence the inverse function of $f(\tilde{\sigma})$, $f^{-1}(\cdot)$, exists. More specifically, we have:

$$\tilde{\sigma} = f^{-1}(L). \tag{10}$$

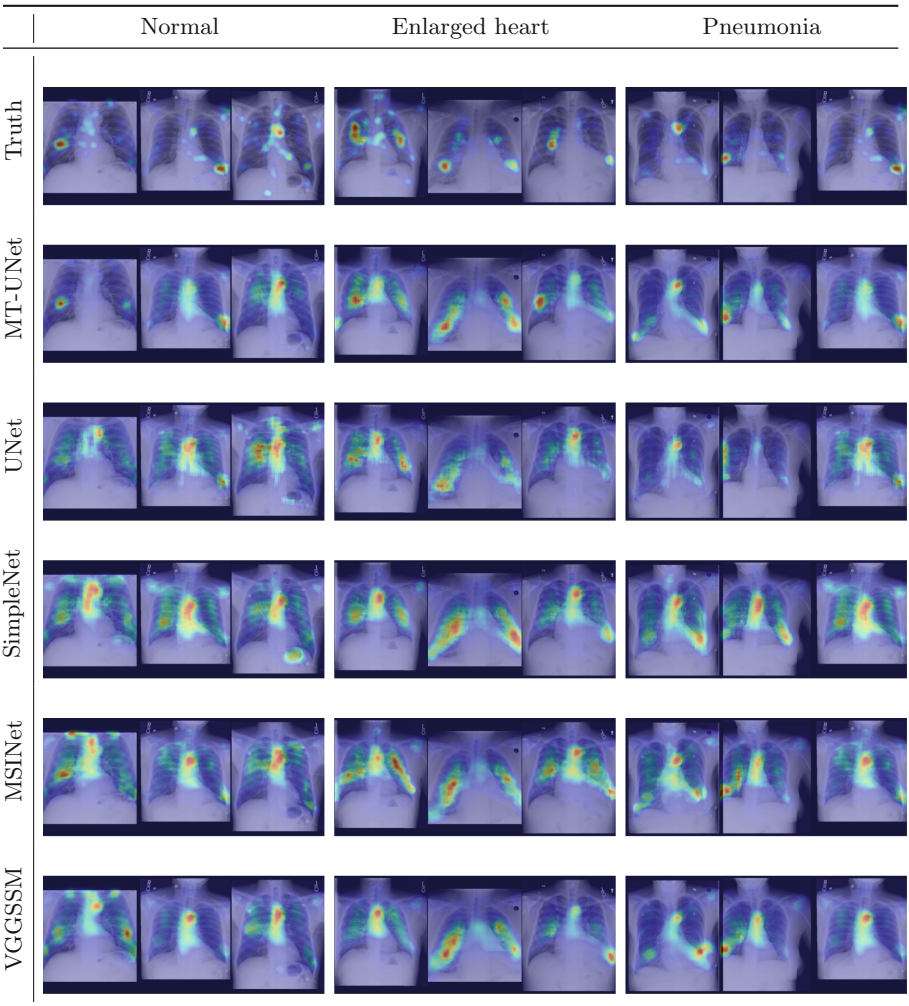
As a pair of inverse functions share the same monotonicity, we know that $\tilde{\sigma} = f^{-1}(L)$ is also strictly monotonically increasing. Thus, when L decreases due to overfitting, we know that $\tilde{\sigma}$ will decrease accordingly, forcing σ to decrease. The decreased σ leads to an increase in the effective learning rate for $\mathcal{T}$, forming a vicious circle of overfitting.

B Training Settings

We use the Adam optimizer with default parameters [29] and the RLRP scheduler for all the training processes. The RLRP scheduler reduces 90% of the learning rate when validation loss stops improving for P consecutive epochs, and reset model parameters to an earlier epoch when the network achieves the best validation loss. All training and testing are performed with the PyTorch framework [45]. Hyper-parameters for optimizations are learning rate r, and P in RLRP scheduler. The dataset is randomly partitioned into 70%, 10% and 20% subsections for training, validation and testing, respectively. The random data partitioning process preserves the balanced dataset characteristic, and all classes have equal share in all sub-datasets. All the results presented in this paper are based on at least 5 independent training with the same hyper-parameters. NVIDIA V100 and A100 GPUs (Santa Clara, USA) were used.

C Saliency Map visualization

Table 4. Visualization of predicted saliency distributions. The ground truth and predicted saliency distributions are overlaid over CXR images. Jet colormap is used for saliency distributions where warmer (red and yellow) colors indicate higher concentration of saliency and colder (green and blue) colors indicate lower concentration of saliency.



References

1. Amyar, A., Modzelewski, R., Li, H., Ruan, S.: Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: classification and segmentation. Comput. Biol. Med. **126**, 104037 (2020)
2. Borji, A.: Saliency prediction in the deep learning era: successes and limitations. IEEE Trans. Patt. Anal. Mach. Intell. **43**, 679–700 (2019)
3. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. IEEE Trans. Image Process. **22**(1), 55–69 (2012)
4. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 740–757 (2018)
5. Cai, Y., Sharma, H., Chatelain, P., Noble, J.A.: Multi-task SonoEyeNet: detection of fetal standardized planes assisted by generated sonographer attention maps. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 871–879. Springer (2018). https://doi.org/10.1007/978-3-030-00928-1_98
6. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K.: Deep learning for chest x-ray analysis: a survey. Med. Image Anal. **72**, 102125 (2021)
7. Cao, G., Tang, Q., Jo, K.: Aggregated deep saliency prediction by self-attention network. In: Huang, D.-S., Premaratne, P. (eds.) ICIC 2020. LNCS (LNAI), vol. 12465, pp. 87–97. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60796-8_8
8. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
9. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nat. Commun. **11**(1), 1–10 (2020)
10. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: International Conference on Machine Learning, pp. 794–803. PMLR (2018)
11. Crawshaw, M.: Multi-task learning with deep neural networks: a survey. arXiv preprint arXiv:2009.09796 (2020)
12. Duffner, S., Garcia, C.: An online backpropagation algorithm with validation error-based adaptive learning rate. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D. (eds.) ICANN 2007. LNCS, vol. 4668, pp. 249–258. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74690-4_26
13. El Asnaoui, K., Chawki, Y., Idri, A.: Automated methods for detection and classification pneumonia based on X-Ray images using deep learning. In: Maleh, Y., Baddi, Y., Alazab, M., Tawalbeh, L., Romdhani, I. (eds.) Artificial Intelligence and Blockchain for Future Cybersecurity Applications. SBD, vol. 90, pp. 257–284. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-74575-2_14
14. Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
15. Fu, K., Dai, W., Zhang, Y., Wang, Z., Yan, M., Sun, X.: MultiCAM: multiple class activation mapping for aircraft recognition in remote sensing images. Remote Sens. **11**(5), 544 (2019)
16. Guo, M., Haque, A., Huang, D.A., Yeung, S., Fei-Fei, L.: Dynamic task prioritization for multitask learning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 270–287 (2018)

17. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learn. **45**(2), 171–186 (2001)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
20. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
21. Jha, A., Kumar, A., Pande, S., Banerjee, B., Chaudhuri, S.: MT-UNET: a novel U-Net based multi-task architecture for visual scene understanding. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 2191–2195. IEEE (2020)
22. Jia, S., Bruce, N.D.: EML-NET: an expandable multi-layer network for saliency prediction. Image Vis. Comput. **95**, 103887 (2020)
23. Johnson, A.E., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 1–8 (2019)
24. Johnson, A.E., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
25. Karargyris, A., et al.: Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. Sci. Data **8**(1), 1–18 (2021)
26. Karessli, N., Akata, Z., Schiele, B., Bulling, A.: Gaze embeddings for zero-shot image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4525–4534 (2017)
27. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491 (2018)
28. Khan, W., Zaki, N., Ali, L.: Intelligent pneumonia identification from chest x-rays: a systematic literature review. IEEE Access **9**, 51747–51771 (2021)
29. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Kroner, A., Senden, M., Driessens, K., Goebel, R.: Contextual encoder-decoder network for visual saliency prediction. Neural Netw. **129**, 261–270 (2020)
31. Kümmerer, M., Wallis, T.S., Bethge, M.: DeepGaze II: reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563 (2016)
32. Li, H., Li, J., Guan, X., Liang, B., Lai, Y., Luo, X.: Research on overfitting of deep learning. In: 2019 15th International Conference on Computational Intelligence and Security (CIS), pp. 78–81. IEEE (2019)
33. Li, Y., Zhang, Z., Dai, C., Dong, Q., Badrigilan, S.: Accuracy of deep learning for automated detection of pneumonia using chest x-ray images: a systematic review and meta-analysis. Comput. Biol. Med. **123**, 103898 (2020)
34. Liebel, L., Körner, M.: Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334 (2018)
35. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
36. Liu, X., Milanova, M.: Visual attention in deep learning: a review. Int. Rob. Auto J. **4**(3), 154–155 (2018)
37. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

38. McLaughlin, L., Bond, R., Hughes, C., McConnell, J., McFadden, S.: Computing eye gaze metrics for the automatic assessment of radiographer performance during x-ray image interpretation. Int. J. Med. Inform. **105**, 11–21 (2017)
39. Moody, G., Mark, R., Goldberger, A.: PhysioNet: a research resource for studies of complex physiologic and biomedical signals. In: Computers in Cardiology 2000, vol. 27 (Cat. 00CH37163), pp. 179–182. IEEE (2000)
40. Moradi, S., et al.: MFP-Unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. Physica Med. **67**, 58–69 (2019)
41. Oyama, T., Yamanaka, T.: Fully convolutional DenseNet for saliency-map prediction. In: 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), pp. 334–339. IEEE (2017)
42. Oyama, T., Yamanaka, T.: Influence of image classification accuracy on saliency map estimation. CAAI Trans. Intell. Technol. **3**(3), 140–152 (2018)
43. Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 598–606 (2016)
44. Paneri, S., Gregoriou, G.G.: Top-down control of visual attention by the prefrontal cortex. functional specialization and long-range interactions. Front. Neurosci. **11**, 545 (2017)
45. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32**, 8026–8037 (2019)
46. Reddy, N., Jain, S., Yarlagadda, P., Gandhi, V.: Tidying deep saliency prediction architectures. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10241–10247. IEEE (2020)
47. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
48. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: why did you say that? arXiv preprint arXiv:1611.07450 (2016)
49. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. arXiv preprint arXiv:1810.04650 (2018)
50. Serte, S., Serener, A., Al-Turjman, F.: Deep learning in medical imaging: a brief review. Trans. Emerg. Telecommun. Technol. 14 (2020)
51. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
52. Smith, L.N.: A disciplined approach to neural network hyper-parameters: part 1-learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
53. Sun, Y., Zhao, M., Hu, K., Fan, S.: Visual saliency prediction using multi-scale attention gated network. Multimedia Syst. **28**(1), 131–139 (2021). https://doi.org/10.1007/s00530-021-00796-4
54. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)
55. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
56. Tan, M., Le, Q.V.: Efficientnetv2: smaller models and faster training. arXiv preprint arXiv:2104.00298 (2021)

57. Tieleman, T., Hinton, G., et al.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. **4**(2), 26–31 (2012)
58. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: a survey. IEEE Trans. Patt. Anal. Mach. Intell. **44**(7) (2021)
59. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12695–12705 (2020)
60. Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., Borji, A.: Revisiting video saliency prediction in the deep learning era. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 220–237 (2019)
61. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)
62. Zhang, Y., Yang, Q.: A survey on multi-task learning. In: IEEE Transactions on Knowledge and Data Engineering (2021). https://doi.org/10.1109/TKDE.2021.3070203
63. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
64. Zhou, Y., et al.: Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. Med. Image Anal. **70**, 101918 (2021)
65. Zhu, H., Salcudean, S., Rohling, R.: Gaze-guided class activation mapping: leveraging human attention for network attention in chest x-rays classification. arXiv preprint arXiv:2202.07107 (2022)
66. Zhu, H., Salcudean, S.E., Rohling, R.N.: A novel gaze-supported multimodal human-computer interaction for ultrasound machines. Int. J. Comput. Assist. Radiol. Surg. **14**(7), 1107–1115 (2019)