Meta Ordinal Weighting Net For Improving Lung Nodule Classification

Yiming Lei[1] Hongming Shan*[2] Junping Zhang[1]

[1]School of Computer Science, Fudan University [2]Institute of Science and Technology for Brain-inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University

ICASSP 2021 TORONTO Canada June 6-11, 2021

Problem: Lung Nodule Classification

Goal: Classifying CT-based lung nodule images (LIDC-IDRI) into three categories: benign ($score \leq 2.5$), unsure ($2.5 < score < 3.5$) and malignant ($score \geq 3.5$).

Motivations:

- Existing methods for lung nodule classification calculate the cross-entropy (CE) loss between the output of the network and the ground-truth label.
- Recently, ordinal regression-based methods take the malignancy progression into consideration, and they have modified the classification head or the label space to improve the classification performance.
- The above two kinds of methods **do not utilize the intrinsic ordinal relationship of data** directly.
- In this paper, we propose to align each training sample with meta ordinal set which contains meta samples of all classes, and the learned meta knowledge guides the learning procedure on training samples.

Method

Meta Ordinal Set (MOS) We assume that the ordinal relationship resides in not only the label but also the data itself. Therefore, we align each target training sample with an MOS that contains K samples from each class. The MOS for i-th training sample is formally defined as follows:

$$\mathcal{S}_i^{\text{meta}} = \left\{\boldsymbol{x}_c^k\right\}_{c=1,\ldots,C,\ k=1,\ldots,K},$$

where C is the number of the classes. Note that the samples in **the $\mathcal{S}^{\text{meta}}$ are not ordered**, but the samples in one class should go to the corresponding multi-layer perceptron (MLP). Then the MPLs are able to learn the specific knowledge from each class. For a target training sample $\boldsymbol{x}_i$, $\boldsymbol{x}_c^k (c \in \{1, 2, \ldots, C\})$ is randomly sampled from the training set, and $\boldsymbol{x}_i \notin \mathcal{S}_i^{meta}$.

Meta Cross-Entropy Loss: To enable the MOW-Net to absorb the meta knowledge provided by the MOS, we propose an MCE loss to align the meta knowledge of each class to the corresponding entropy term: $\mathcal{L}_{\text{MCE}} = -\sum_{c=1}^{C} \omega_c \log \hat{p}_c$, where $\hat{p}_c$ and ω_c are the prediction and the learned meta weight of the c-th class, respectively. Note that the MCE loss implies no ordinal regression tricks such as cumulative probabilities, it only holds the correlation between the meta data and the predictions. Compared with the conventional CE loss, **the MCE loss enables the training samples to be supervised by the corresponding meta data**, hence, the learning of the MOW-Net takes into account the meta ordinal knowledge resided in the data itself.

Training Objectives:

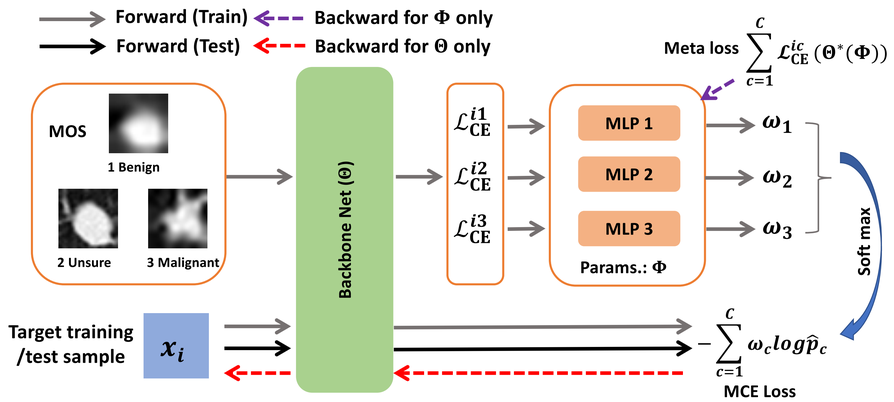
$$\boldsymbol{\Theta}^*(\boldsymbol{\Phi}) = \arg\min_{\boldsymbol{\Theta}} \mathcal{L}_{\text{MCE}}(\boldsymbol{\Theta}; \boldsymbol{\Phi}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{MCE}}^i = \frac{-1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \underbrace{V_c(\mathcal{L}_{\text{CE}}^{i,c}(\boldsymbol{\Theta}); \boldsymbol{\Phi})}_{\omega_c} \cdot \log \hat{p}_{i,c}(\boldsymbol{\Theta})$$

$$\boldsymbol{\Phi}^* = \arg\min_{\boldsymbol{\Phi}} \frac{1}{M}\sum_{j=1}^{M}\sum_{c=1}^{C} \mathcal{L}_{\text{CE}}^{j,c}(\boldsymbol{\Theta}^*(\boldsymbol{\Phi})),$$

$$\boldsymbol{\Phi}^{(t+1)} = \boldsymbol{\Phi}^{(t)} + \frac{\alpha\beta}{N}\sum_{i=1}^{N}\left(\frac{1}{M}\sum_{j=1}^{M} G_{ij}\right) \frac{\partial \sum_{c=1}^{C} V_c(\mathcal{L}_{\text{CE}}^{i,c}(\boldsymbol{\Theta}^{(t)}); \boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}}\Big|_{\boldsymbol{\Phi}^{(t)}},$$

Meta Ordinal Weighting Net (MOW-Net) & Experiments

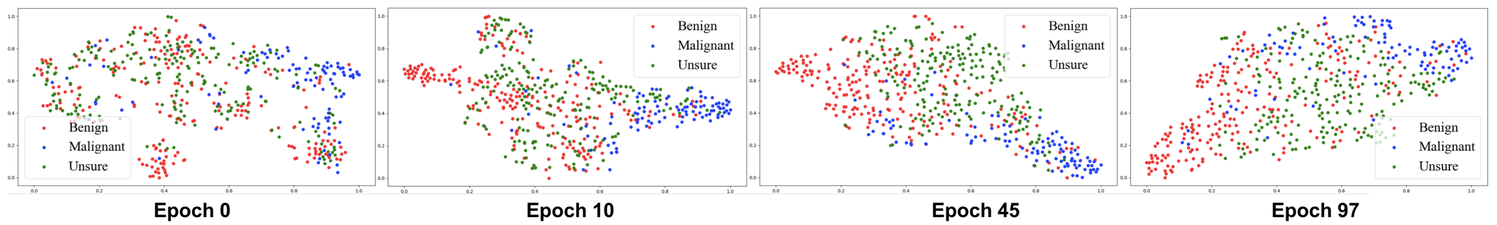
The MOW-Net framework:



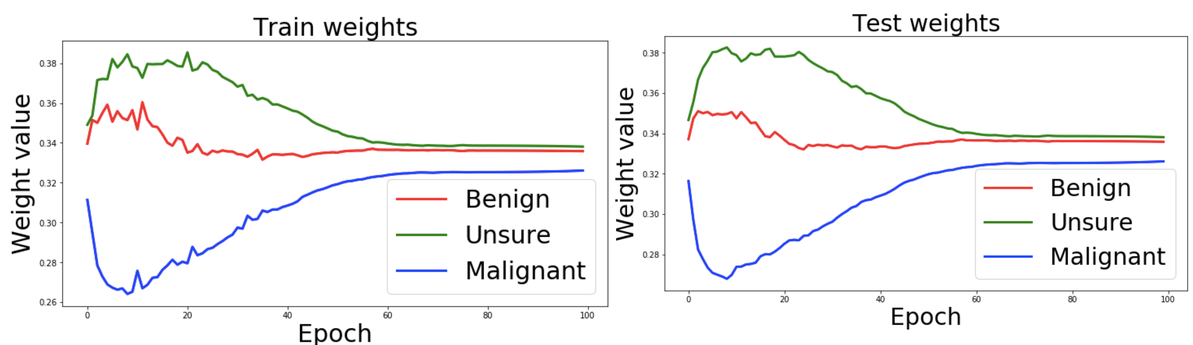
Classification performance:

Method	Accuracy	Benign			Malignant			Unsure		
		P	R	F1	P	R	F1	P	R	F1
CE Loss	0.517	0.538	0.668	0.596	0.562	0.495	0.526	0.456	0.360	0.402
Poisson (ICML 2017)	0.542	0.548	<u>0.794</u>	0.648	0.568	0.624	0.594	0.489	0.220	0.303
NSB (ECCV 2018)	0.553	0.565	0.641	0.601	0.566	0.594	0.580	0.527	0.435	0.476
UDM (ICCV 2019)	0.548	0.541	0.767	0.635	<u>0.712</u>	0.515	0.598	0.474	0.320	0.382
CORF (TNNLS 2021)	0.559	0.590	0.627	0.608	0.704	0.495	0.581	0.476	0.515	0.495
MOW-Net ($k = 1$)	0.629	0.752	0.489	0.592	0.558	**0.851**	0.675	0.600	0.675	0.635
MOW-Net ($k = 5$)	0.672	0.764	0.596	0.670	0.600	0.802	0.686	<u>0.642</u>	0.690	0.665
MOW-Net ($k = 10$)	**0.687**	**0.768**	0.623	**0.688**	0.668	0.705	**0.686**	0.606	**0.792**	**0.687**

Feature visualizations:



Variations of learned weights:



Training & inference: **Note that the MOS only works in the training stage. The final goal of the optimization is to obtain the optimal Θ^*.** From above equations, we have:

$$G_{i,j} = \left[\frac{\partial \sum_{c=1}^{C} \mathcal{L}_{\text{CE}}^{j,c}(\widehat{\boldsymbol{\Theta}})}{\partial \widehat{\boldsymbol{\Theta}}}\right]^{\mathrm{T}} \cdot \frac{\partial \sum_{c=1}^{C} \log \hat{p}_{i,c}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}.$$

$G_{i,j}$ represents the similarity between the training data and the MOS data, which guarantees the gradients of them to have identical direction.