Self-Supervised, Semi-Supervised, Multi-Context Learning for the Combined Classification and Segmentation of Medical Images (Student Abstract)

Abdullah-Al-Zubaer Imran,[1,2] **Chao Huang,**[2] **Hui Tang,**[2] **Wei Fan,**[2]
Yuan Xiao,[3] **Dingjun Hao,**[3] **Zhen Qian,**[2] **Demetri Terzopoulos**[1,4]
[1]UCLA Computer Science Department, Los Angeles, California USA
[2]Tencent Medical AI Lab, Palo Alto, California, USA
[3]Xi'an Jiaotong University College of Medicine, China
[4]VoxelCloud, Inc., Los Angeles, California, USA

Abstract

To tackle the problem of limited annotated data, semi-supervised learning is attracting attention as an alternative to fully supervised models. Moreover, optimizing a multiple-task model to learn "multiple contexts" can provide better generalizability compared to single-task models. We propose a novel semi-supervised multiple-task model leveraging self-supervision and adversarial training—namely, self-supervised, semi-supervised, multi-context learning (S^{4}MCL)—and apply it to two crucial medical imaging tasks, classification and segmentation. Our experiments on spine X-rays reveal that the S^{4}MCL model significantly outperforms semi-supervised single-task, semi-supervised multi-context, and fully-supervised single-task models, even with a 50% reduction of classification and segmentation labels.

Introduction

Depending on how unlabeled data are leveraged, semi-supervised learning can be accomplished in several ways, and this has recently emerged as a growing body of research, yielding schemes such as unsupervised domain adaptation (Zhang et al. 2019), self-supervised learning (Jing and Tian 2019), adversarial learning (Imran and Terzopoulos 2019a), and multi-task learning (Ruder 2017). We propose a self-supervised, semi-supervised, multi-context learning (S^{4}MCL) model that combines the advantages of self-supervised learning, adversarial learning, and multi-task learning for use in real-world applications. To demonstrate its effectiveness, we apply our model to two of the most important tasks in medical imaging—disease classification and segmentation of anatomical structures—and both tasks are tackled by the same model, thus satisfying the clinical need to label an image as normal or abnormal as well as to segment the relevant anatomical structures that are imaged.

Methods

To formulate the problem, we assume an unknown data distribution $p(X, Y, Z)$ over images X, segmentation labels Y, and class labels Z. The model has access to the labeled training set $\mathcal{D}_\mathcal{L}$ sampled i.i.d. from $p(X, Y, Z)$ and unlabeled training set $\mathcal{D}_\mathcal{U}$ sampled i.i.d. from $p(X)$ after marginalizing out Y and Z. Two networks, S and C, are trained in an adversarial learning fashion, such that the mask generator S and the class discriminator C compete against each other. We specify the objective as two losses: $\min_{\phi_C} \mathcal{L}_C(\mathcal{L}(\mathcal{D}_\mathcal{L}, \phi_C) + \alpha\mathcal{L}(\mathcal{D}_\mathcal{U}, \phi_C))$ and $\min_{\phi_S} \mathcal{L}_S(\mathcal{L}(\mathcal{D}_\mathcal{L}, \phi_S) + \alpha\mathcal{L}(\mathcal{D}_\mathcal{U}, \phi_S))$, where ϕ_C and ϕ_S are the parameters of networks C and S, respectively.

For the classification, we use a transformation function $t(x)$ to make random flipping (horizontal/vertical) or rotation (0, 90, 180, etc.) for the unlabeled images and let the network C predict them. S's supervised loss is just based on the labeled samples (at pixel-level). We employ the generalized Dice loss in this regard. L_S includes segmentation loss and adversarial prediction loss. Since the main objective

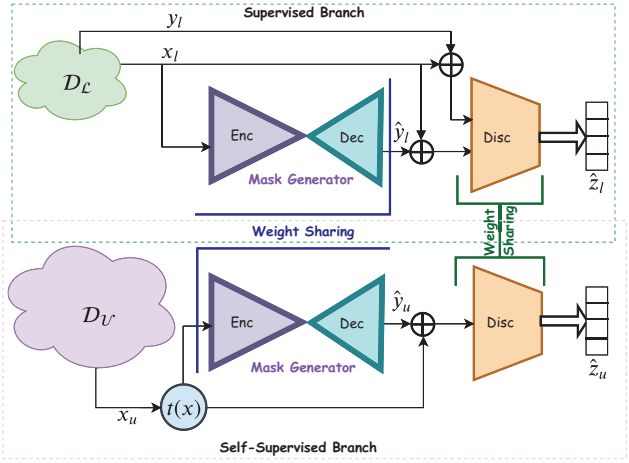

Figure 1: Our S^{4}MCL model. A segmentation mask generator produces masks from inputs of labeled or unlabeled samples. A discriminator takes concatenated inputs of labeled data-mask or unlabeled data-mask pairs and predicts the class labels. With the labeled data branch, it is supervised. By contrast, unlabeled data is self-supervised based on self-generated labels using a transformation function $t(x)$. Segmentation output is obtained from the decoder (Dec) and classification output is received at the discriminator (Disc).

Table 1: Performance comparison of our S^{4}MCL model against the baselines in different data settings with varying percentages of labeled data: accuracy (Acc) for classification and Dice score (DS) for segmentation have been reported.

Single-Task			Multi-Task					
Model	Acc	DS	Model	Acc	DS	Model	Acc	DS
CNN-5	0.420	0.702	S^{2}MCL-5	0.420	0.500	S^{4}MCL-5	0.630	0.890
CNN-10	0.450	0.874	S^{2}MCL-10	0.460	0.640	S^{4}MCL-10	0.670	0.907
CNN-20	0.460	0.903	S^{2}MCL-20	0.500	0.672	S^{4}MCL-20	0.680	0.921
CNN-30	0.500	0.910	S^{2}MCL-30	0.550	0.752	S^{4}MCL-30	0.740	0.925
CNN-50	0.620	0.919	S^{2}MCL-50	0.670	0.889	S^{4}MCL-50	**0.800**	**0.934**
CNN-100	0.780	0.931						

of S is to generate the segmentation map, a small weight (0.01) is used for the adversarial loss terms. C is trained on multiple objectives—adversary on the segmentation mask generator S's output and classification of the images into the real or surrogate classes. For the labeled examples, we calculate two-way losses from image-label and image-prediction pairs, which differs from the unlabeled examples, where only image-prediction pairs are taken into account. The unsupervised adversarial loss terms include adversarial losses for the labeled and unlabeled data. L_C includes supervised classification loss on $(\hat{z}_l|x_l, y_l)$, self-supervised classification loss on $(\hat{z}_u|x_u, \hat{y}_u)$, adversarial real loss on (x_l, y_l), and adversarial prediction losses on $(x_l, \hat{y}_l)$ and $(x_u, \hat{y}_u)$.

Experimental Evaluation

We use 894 patches extracted from 100 spinal X-Rays with vertebrae masks obtained from a vertebral compression fracture study of osteoporosis (Wong and McGirt 2013). The patches are split into three subsets: training set (713), validation set (42), and testing set (139). our S^{4}MCL is compared against baseline models—semi-supervised multi-context learning (S^{2}MCL), conv-net (U-Net) for segmentation (Ronneberger, Fischer, and Brox 2015), and conv-net for classification. All the images are normalized and resized to $128 \times 128 \times 1$ before feeding them to the models. We use a U-Net-like encoder-decoder network with skip connections as the segmentation mask generator and another convolutional network as the class discriminator (Imran and Terzopoulos 2019b).

Our S^{4}MCL model performs better than all the baseline semi-supervised and fully-supervised models both in the segmentation of vertebrae and predicting the fractured ones (Table 1). Its consistently good performance with a varying proportion of labeled training data confirms the robustness of our S^{4}MCL model. Visualization of segmented vertebra boundaries by different models under various labeled data proportions (Fig. 2) reveals the superiority of our model.

Note that for a fair comparison of all the models, we use a common segmentation architecture in single-task for segmentation and in multi-task for segmentation mask generator. Our S^{4}MCL model proves to have a consistently better performance than the semi-supervised and fully-supervised single-task segmentation (U-Net) model, given the same proportion of labeled training data ($|\mathcal{D}_\mathcal{L}|$). The advantage really accrues with the knowledge gain from the larger portion of unlabeled data and multi-context learning.

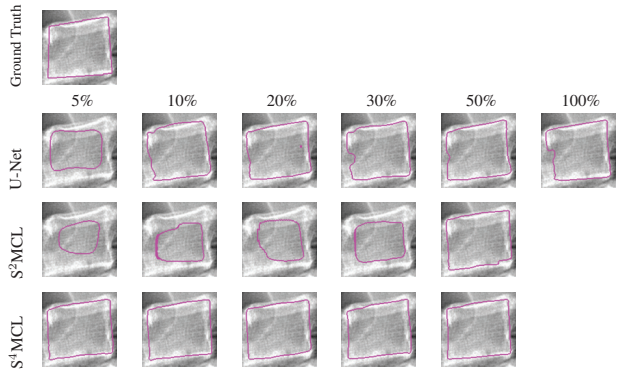

Figure 2: Boundary visualization of a predicted vertebra mask showing the superiority of our S^{4}MCL model with a varying proportion of labeled data.

Conclusions

Learning from small labeled datasets have been one of the most challenging tasks in computer vision and medical imaging. We proposed a novel self-supervised, semi-supervised, multi-context learning (S^{4}MCL) model, which we validated through medical image classification and segmentation experiments with limited labeled data. Our experimental results confirmed the superiority of our S^{4}MCL model over semi-supervised and fully-supervised single-context and multi-context models.

References

Imran, A.-A.-Z., and Terzopoulos, D. 2019a. Multi-adversarial variational autoencoder networks. In *IEEE International Conference on Machine Learning and Applications*, 777–782.

Imran, A.-A.-Z., and Terzopoulos, D. 2019b. Semi-supervised multi-task learning with chest X-Ray images. In *International Workshop on Machine Learning in Medical Imaging*, 151–159. Cham: Springer.

Jing, L., and Tian, Y. 2019. Self-supervised visual feature learning with deep neural networks: A survey. *arXiv preprint. arXiv:1902.06162 [cs.CV]. Ithaca, NY: Cornell University Library*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 234–241. Cham: Springer.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint. arXiv:1706.05098 [cs.LG]. Ithaca, NY: Cornell University Library*.

Wong, C. C., and McGirt, M. J. 2013. Vertebral compression fractures: a review of current management and multimodal therapy. *Journal of Multidisciplinary Healthcare* 6:205.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, 7404–7413.