Multiview Long-Short Spatial Contrastive Learning for 3D Medical Image Analysis

Gongpeng Cao[1], Yiping Wang[1], Manli Zhang[1], Jing Zhang[1], Guixia Kang[1], Xin Xu[2]

[1]Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China

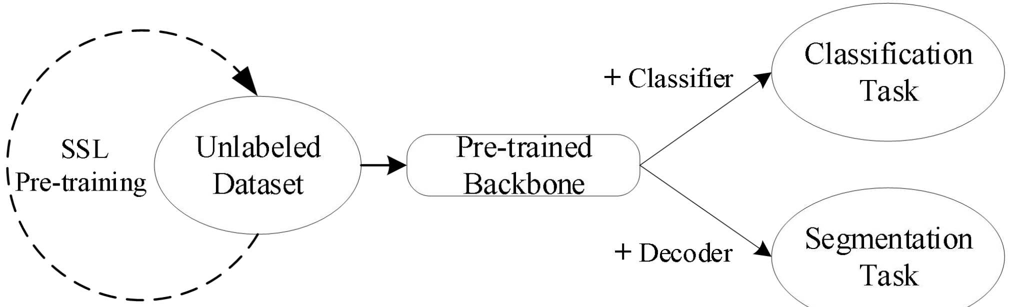
[2]Department of Neurosurgery, General Hospital of PLA, Beijing 100853, China

ICASSP 2022

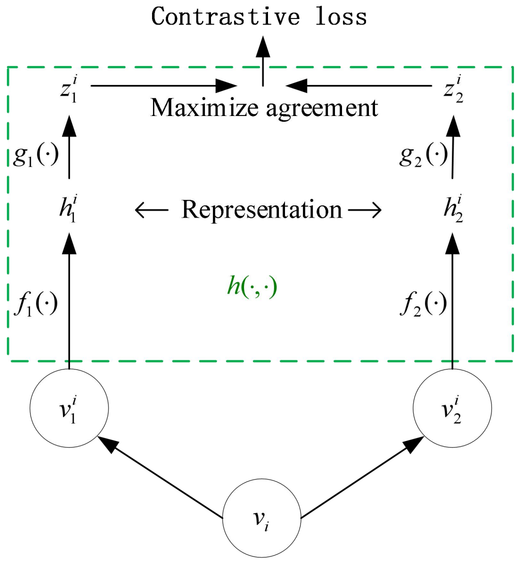
Introduction

Background:

- Supervised deep learning requires sufficient labeled data.
  - Expert knowledge in specific fields.
  - Time-consuming and Laborious.
- Self-supervised learning (SSL):
  - Without human-annotations.
  - Generic representations transferred to downstream tasks.



Contrastive Learning:



- An effective implementation of the self-supervised learning.
- Contrastive learning trains neural networks to discriminate between "positive" pairs (v_1^i, v_2^i) and "negative" pairs $(v_1^i, v_2^j)_{j\neq i}$.
- Learning is formulated as minimizing a contrastive loss (we use InfoNCE in this paper).

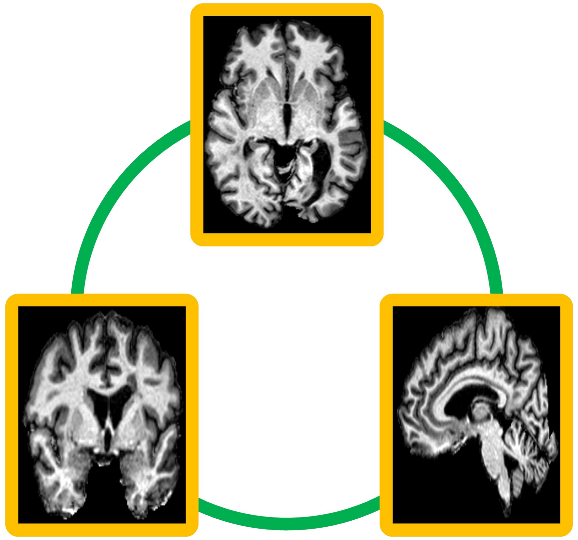
$$L_N(v_1, v_2) = -\mathbb{E}\left[log \frac{e^{h(v_1^i, v_2^i)}}{\sum_{j=1}^{K+1} e^{h(v_1^i, v_2^j)}}\right]$$

Motivation

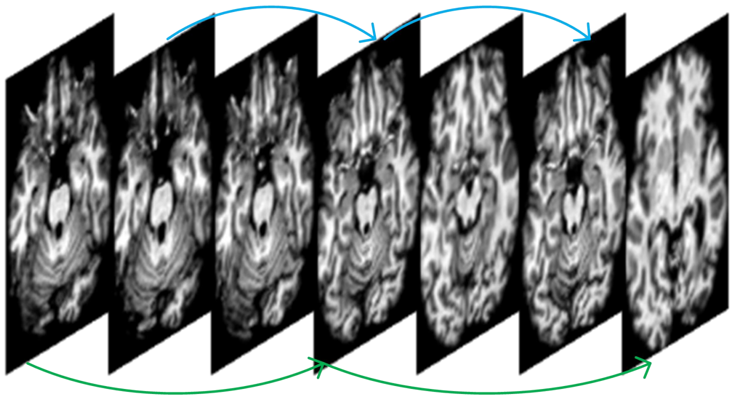
Two limitations of existing contrasting strategies when applied to 3D medical images:

- ➢ Ignore the intrinsic structural similarity.
- ➢ Ignore local representation.

Observation 1：
The information shared between three views (axial, coronal and sagittal views) can capture the global representation of volumetric medical image.



Observation 2：
Matching the short spatial clip to long spatial clip forces the model to extrapolate local information.



Method: Multiview Contrasting Strategy & Long-Short Spatial Contrasting Strategy

Multiview Contrasting Strategy:

- To learn global representation, we need to maximize the mutual information between three views(v_a, v_c, v_s)

$$\max\{I(v_a; v_c) + I(v_a; v_s) + I(v_c; v_s)\}$$

- InfoNCE loss can estimate the lower bound of mutual information. For two views v_1, v_2:

$$I(v_1; v_2) \geq \log(K) - L_N(v_1, v_2)$$

- Maximizing mutual information between three views → Multiview contrastive learning:

$$L_{multiview} = L_N(v_a, v_c) + L_N(v_a, v_s) + L_N(v_c, v_s)$$

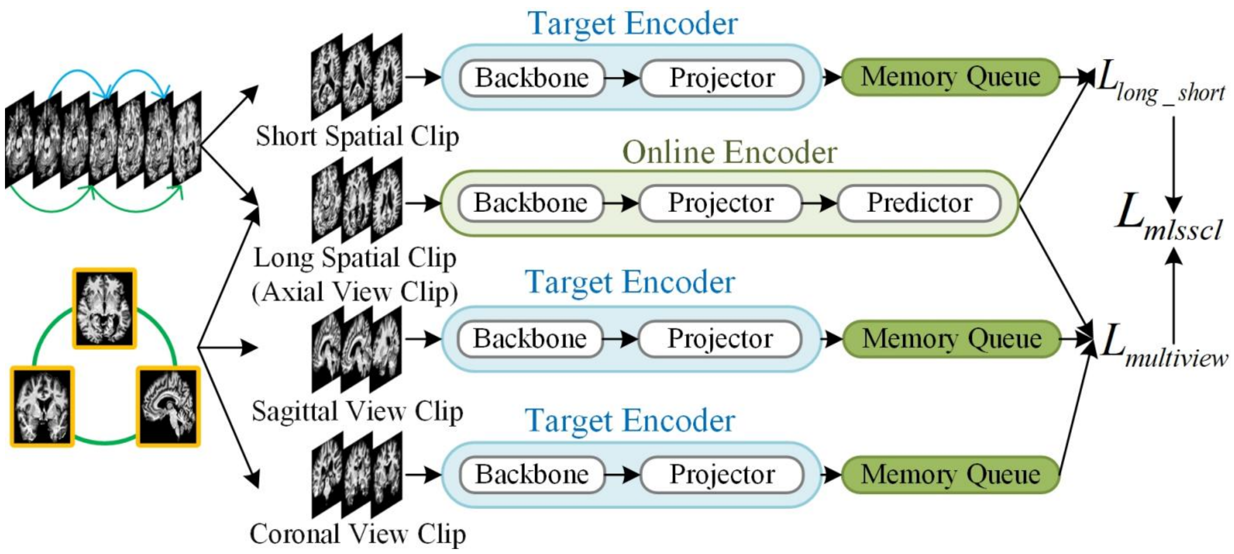
Long-Short Spatial Contrasting Strategy:

- Maximizing representation similarity between a long spatial clip v_L and a much shorter spatial clip v_S:

$$L_{long-short} = L_N(v_S, v_L)$$

Matching the short-clip representation to the long-clip representation forces the model to understand and recognize the structure and correlation of local tissues in volumetric medical images.

Method: Multiview Long-Short Spatial Contrastive Learning Framework



Clip Sampling:

- Sample axial, coronal and sagittal clips from a 3D volumetric medical image with C slices and a stride of δ_L.
- Regard the above axial clip as the long spatial clip and then randomly sample C axial slices with spatial stride $\delta_S(\delta_S < \delta_L)$ as the short spatial clip.

Network Architecture:

- One online encoder:
  - a backbone + a projector head (2-layer MLP) + a prediction head (2-layer MLP).
  - updated by back-propagation.
  - the backbone will be transferred to downstream tasks after pre-training.
- Three target encoders: (share weights)
  - a backbone + a projector head (2-layer MLP)
  - updated in the manner of momentum.
  - memory queue to store previous representations.
  - discarded after pre-training.

Contrastive Loss:

$$L_{mlsscl} = \alpha L_{multiview} + \beta L_{long\_short}$$

Experiments: Pre-training on Large-Scale Unlabeled Dataset

Pre-training Dataset:

- ADNI pre-training set (5953 T1-weighted MRI scans).

Instantiation of Network:

- AD classification task:
  - 3D ResNet-18 as backbone.
- MS lesion segmentation task:
  - 3D UNet-based encoder as backbone.

Optimization:

- We pre-train models on ADNI pre-training set for 100 epochs with SGD optimizer.

Other details can be found in paper.

Experiments: Transferring Learned Features to AD Classification

Table 1. Results(mean±std) for AD classification (AD vs. HC) on the ADNI-AD classification test set.

Method	ACC	SEN	SPE	AUC
Training-from-Scratch	0.793 ± 0.011	0.874 ± 0.055	0.711 ± 0.058	0.896 ± 0.006
BYOL [12]	0.809 ± 0.004	0.866 ± 0.032	0.752 ± 0.037	0.886 ± 0.016
MoCo [10]	0.825 ± 0.020	0.886 ± 0.043	0.764 ± 0.060	0.895 ± 0.001
Model Genesis [8]	0.827 ± 0.004	0.911 ± 0.061	0.744 ± 0.061	0.904 ± 0.009
Age-Aware [13]	0.831 ± 0.007	0.882 ± 0.007	0.780 ± 0.012	0.899 ± 0.011
MLSSCL	$\mathbf{0.858 \pm 0.013}$	$\mathbf{0.911 \pm 0.019}$	$\mathbf{0.805 \pm 0.044}$	$\mathbf{0.907 \pm 0.012}$

- MLSSCL achieves a remarkable improvement over training-from-scratch: ↑ 6.5%(ACC), ↑ 3.7%(SEN), ↑ 9.4%(SPE)
- MLSSCL outperforms other SSL methods: ↑ 2.7%(ACC), ↑ 2.5%(SPE)

- MLSSCL can effectively deal with the situation with few labeled training samples. 70% labeled data (MLSSCL) ≈ 100% labeled data (training-from-scratch)

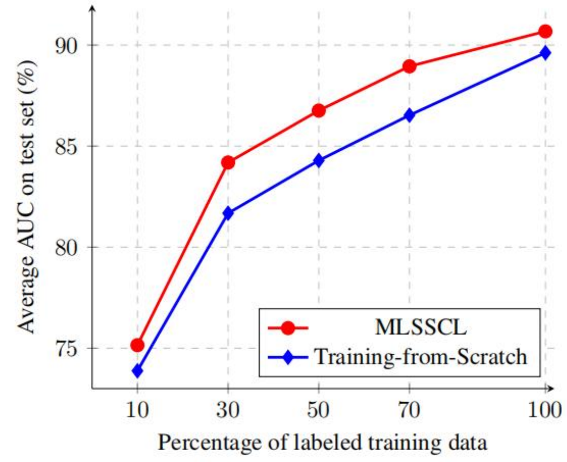


Fig. 2. The AD classification performance of networks trained with different amounts of labeled data.

Experiments: Transferring Learned Features to MS Lesion Segmentation

Table 2. The segmentation results of different approaches on the ISBI 2015 longitudinal MS lesion segmentation test set.

Method	DSC†	PPV†	LTPR†	LFPR†
Training-from-Scratch	0.6176	0.8229	0.4451	0.3485
SSL				
Age-Aware [13]	0.6320	0.8103	0.4586	0.3034
BYOL [12]	0.6337	0.7991	0.4675	0.3442
MoCo [10]	0.6369	0.7972	0.4641	0.3092
Model Genesis [8]	0.6434	0.8200	0.4647	0.3082
MS SOTA				
Aslania et al. [3]	0.6114	**0.8992**	0.4103	0.1393
Andermatt et al. [18]	0.6298	0.8446	0.4870	0.2013
Valverde et al. [4]	0.6304	0.7866	0.3669	0.1529
Hu et al. [5]	0.6345	0.8682	0.4787	**0.1299**
MLSSCL	**0.6482**	0.8007	**0.4933**	0.2796

- MLSSCL consistently outperforms training-from-scratch and other SSL methods. Compared with training-from-scratch:
  ↑ 3.06%(DSC), ↑ 4.82%(LTPR), ↑ 6.89%(LFPR)
- MLSSCL still achieves higher DSC and LTPR compared to SOTA segmentation methods.

Experiments: Ablation to contrasting strategies on AD classification

Table 3. Ablation to contrasting strategies on AD classification task (mean ± std).

Contrasting Strategy	**ACC**	**AUC**
Long-Short	0.823 ± 0.012	0.892 ± 0.014
Multiview	0.833 ± 0.027	0.906 ± 0.024
Multiview & Long-Short	$\mathbf{0.858 \pm 0.013}$	$\mathbf{0.907 \pm 0.012}$

The results demonstrate the complementarity of global representation and local representation.

Conclusion

- ✓ We introduce multiview contrasting strategy to learn global representations by maximizing the mutual information between three views of the same volumetric medical image.
- ✓ We introduce long-short spatial contrasting strategy to learn local representations by matching a short spatial clip to a long spatial clip in the latent space under the given view.
- ✓ We propose multiview long-short spatial contrastive learning (MLSSCL) framework to combine these two contrasting strategies, which can effectively learn generic 3D representations.
- ✓ Extensive experimental results showed that MLSSCL outperformed training-from-scratch method, especially when fine-tuned on only small amounts of labeled data, and also showed a clear superiority compared with other self-supervised learning methods.

Thank you!

E-mail:

gpcao@bupt.edu.cn

gxkang@bupt.edu.cn

This work was supported by Fundamental Research Funds for the Central Universities (2020XD-A06-1), the State Key Program of the National Natural Science Foundation of China (82030037), the National Science and Technology Major Project of China (No.2017ZX03001022).