Mutual Contrastive Low-rank Learning to Disentangle Whole Slide Image Representations for Glioma Grading

Lipei Zhang[1], Yiran Wei[2], Ying Fu[3], Stephen Price[2], Carola-Bibiane Schönlieb[1], Chao Li[1,2,*]

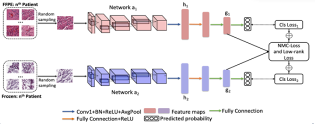
1. Department of Applied Mathematics and Theoretical Physics, University of Cambridge
2. Department of Clinical Neuroscience, University of Cambridge
3. School of Computer Science and Technology, Beijing Institute of Technology

Abstract

- Whole slide images are mainly derived from formalin-fixed paraffin-embedded (FFPE) and Frozen section.
- FFPE sections could be affected by the artifacts introduced from tissue processing. The frozen sections have low quality.
- To overcome these problems in a single modal training and achieve better multi-modal and discriminative representation disentanglement in brain tumor, we propose a mutual contrastive low-rank learning (MCL) scheme which combine mutual learning scheme with a normalized modality contrastive loss (NMC-loss) and a low-rank (LR) loss.
- Proposed scheme achieves better performance on the Cancer Genome Atlas (TCGA) brain dataset than the model trained based on each single modality or mixed modalities and improves the feature extraction in classical attention-based multiple instances learning methods (MIL). The combination of NMC-loss and low-rank loss outperforms other typical contrastive loss functions.

Method

MCL



Non-linear representation

- Adopt three fully connections with ReLU in our scheme to promote better non-linear projection of latent vectors for representation learning.
- Obtain a representative vector with the non-linear projection
- Remove variant information, e.g., the color or orientation of objects resulting from various staining procedures from multiple centers.

$$z_i = g(h_i) = W^2\sigma(W^1 h_i)$$

NMC-loss

- Given a batch size N, FFPE (x^a_1, ..., x^a_k) and the frozen section (x^b_1, ..., x^b_k) images as images sampled from different augmented views on the same patient.
- Adopt layer normalization to re-center and rescale the latent space from two different domain into the same sphere. where n ∈ {1, 2} and g_n denotes the latent vector from FFPE or frozen section. μ and σ are the mean and variance of each batch.

$$\hat{g}_n = \frac{g_n - \mu_n}{\sqrt{(\sigma_n)^2 + \varepsilon}}$$

- The NMC-loss function can be defined as:

$$L_k^a = -log\frac{exp(sim(\hat{g}_k^a, \hat{g}_k^b)/\tau)}{\sum_{i\in I}\mathbb{1}_{i\neq k}exp(sim(\hat{g}_k^a, \hat{g}_i^b)/\tau)} \qquad L_{nmc} = \frac{1}{2N}\sum_{k=1}^{K}(L_k^a + L_k^b)$$

Low-rank loss

- Consider a feature embedding of each modality $X_a = [x^a_1 | ... | x^a_N]$, where each column $x^a_i \in R^d$, i = 1, ..., N, a ∈ [1, 2] and | represents vertical concatenation. We further obtain $M = [X^1 | X^2]$. C denotes category number. The low-rank loss will be defined as:

$$L_{lr} = \sum_{c=1}^{C} max(\Delta, ||M_c||_*) - ||M||_*$$
$$= \sum_{c=1}^{C} max(\Delta, ||[\phi(Y_c^1;\theta)|\phi(Y_c^2;\theta)]||_*) - ||[\phi(Y^1;\theta)|\phi(Y^2;\theta)]||_*$$

Where ||.||* means the matrix nuclear norm (the sum of the singular values) Δ ∈ R denotes a bound on the intra-class nuclear loss that can avoid the training collapse resulting from feature value to zero. In our experiments, we set Δ =1.

Model optimization

- Each function consists of a cross-entropy (CE) loss with a Taylor Softmax loss, and an NMC-loss and a low rank loss.

$$L_{cls}(f(x), y) = \sum_{i=1}^{t}\frac{(1 - f_y(x))^i}{i}$$

$$L_{FFPE} = L_{cls_1} + L_{nmc} + L_{lr}$$
$$L_{Frozen} = L_{cls_2} + L_{nmc} + L_{lr}$$

Where the $f_y(x)$ denotes the y-th element of f(x) and f(.) is a CNN with the classification layer. t is the term number of the Taylor series.

Results

	FFPE			Frozen		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Single training	0.71	0.72	0.71	0.70	0.71	0.70
Mixed training	0.70	0.72	0.71	0.68	0.67	0.68
MCL	**0.76**	**0.77**	**0.76**	**0.74**	**0.73**	**0.74**
ImageNet + A-MIL	0.72	0.72	0.72	0.70	0.68	0.70
ImageNet + TransMIL	0.71	0.72	0.71	0.74	0.73	0.74
ImageNet + CLAM	0.71	0.72	0.71	0.69	0.70	0.69
single training+A-MIL	0.74	0.75	0.74	0.74	0.73	0.74
single training+TransMIL	0.69	0.70	0.73	0.73	0.72	0.73
single training+CLAM	0.75	0.76	0.75	0.73	0.72	0.73
Mixed training+A-MIL	0.73	0.73	0.73	0.70	0.69	0.70
Mixed training+TransMIL	0.70	0.70	0.70	0.72	0.73	0.72
Mixed training+CLAM	0.75	0.75	0.75	0.71	0.69	0.71
MCL+A-MIL	0.78	0.79	0.78	0.75	0.74	0.75
MCL+TransMIL	0.77	0.77	0.77	0.74	0.75	0.74
MCL+CLAM	**0.79**	**0.80**	**0.79**	**0.75**	**0.75**	**0.75**

Table1. Comparison with different learning schemes.

	FFPE			Frozen		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KL-loss	0.71	0.70	0.68	0.71	0.68	0.69
Marginal triplet loss	0.74	0.74	0.74	0.72	0.72	0.72
NT-Logistic loss	0.74	0.74	0.72	0.71	0.68	0.70
NT-Xent loss	0.71	0.71	0.71	0.73	0.73	0.73
AMC loss	0.75	0.76	0.75	0.70	0.70	0.70
LR loss	0.75	0.77	0.75	0.72	0.73	0.72
NMC-loss	0.75	0.75	0.75	0.72	0.71	0.72
NMC-loss + LR Loss	**0.76**	**0.77**	**0.76**	**0.74**	**0.73**	**0.74**

Table II. Comparison with different contrastive loss functions

Visualization

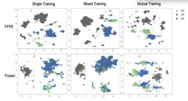


Figure I. Visualization of reducing the dimension of the latent vector by UMAP method.

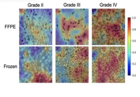


Figure II. Image-level predicted heatmap of three tumor grades by mutual learning

Conclusion

- The MCL training scheme achieve improvement grading performance on the opened TCGA brain WSI dataset.
- Better ability on disentangling discriminative representations.
- The MCL is robust and convincing for boosting the performance of classical attention-based MIL.