Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Computational Biology Department

Open-set Recognition of Unseen Macromolecules in Cellular Electron Cryo-Tomograms by Soft Large Margin Centralized Cosine Loss

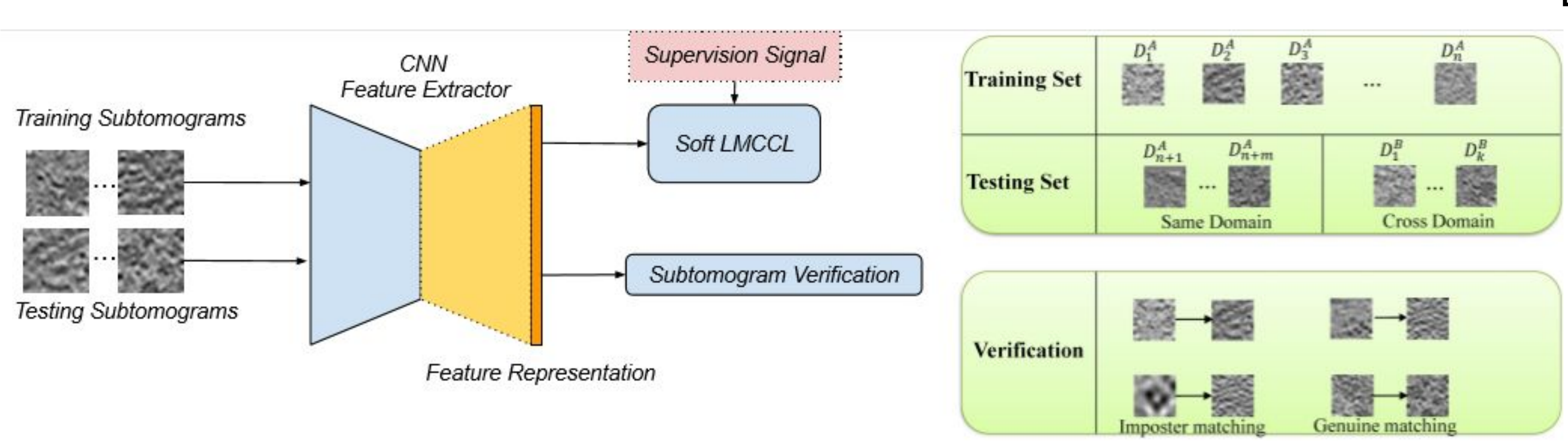
Xuefeng Du[1], Xiangrui Zeng[1], Bo Zhou[1], Alex Singh[1], Min Xu[1]

[1]*Carnegie Mellon University*

Abstract

Cellular Electron Cryo-Tomography (CECT) is a 3D imaging tool that visualizes the structure and spatial organization of macromolecules at sub-molecular resolution in a near native state, allowing systematic analysis of seen and unseen macromolecules. Methods for high-throughput subtomogram classification on known macromolecules based on deep learning have been developed. However, the learned features guided by either the regular Softmax loss or traditional feature descriptors are not well applicable in the open-set recognition scenarios where the testing data and the training data have a different label space. In other words, the testing data contain novel structural classes unseen in the training data. In this paper, we propose a novel loss function for deep neural networks to extract discriminative features for unseen macromolecular structure recognition in CECT, called Soft Large Margin Centralized Cosine Loss (Soft LMCCL). Our Soft LMCCL projects 3D images into a normalized hypersphere that generates features with a large inter-class variance and a low intra-class variance, which can better generalize across data with different classes and in different datasets. Our experiments on CECT subtomogram recognition tasks using both simulation data and real data demonstrate that we are able to achieve significantly better verification accuracy and reliability compared to classic loss functions. In summary, our Soft LMCCL is a useful design in our detection task of unseen structures and is potentially useful in other similar open-set scenarios.

Method



(a) (b)

Fig. 1. (a) The flowchart of open-set macromolecule recognition. (b) The configuration of our training and evaluation protocols.

Softmax Loss

$$L_{softmax} = \frac{1}{N}\sum_{i=1}^{N} -\log p_i = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{W_{y_i}^T \cdot x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T \cdot x_i + b_j}} = \sum_{i=1}^{N} -\log \frac{e^{\|W_{y_i}\|\|x_i\|\cos\theta_{y_i} + b_{y_i}}}{\sum_{j=1}^{C} e^{\|W_j\|\|x_i\|\cos\theta_j + b_j}}$$

LMCL for Inter-Class Variance Maximization

$$L_{softmax} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{s\cdot\cos\theta_{y_i} + b_{y_i}}}{\sum_{j=1}^{C} e^{s\cdot\cos\theta_j + b_j}}. \xrightarrow{\text{Cosine Margin}} \begin{array}{l} C_1 : \cos\theta_1 > \cos\theta_2 + m \\ C_2 : \cos\theta_2 > \cos\theta_1 + m. \end{array}$$

$$L_{LMCL} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\left(\cos\theta_{y_i} - m\right) + b_{y_i}}}{e^{s\left(\cos\theta_{y_i} - m\right) + b_j} + \sum_{j=1, j\neq y_i}^{C} e^{s\cos\theta_j + b_j}}$$

Center Loss for Intra-class Minimization

$$L_{Center} = \frac{1}{2}\sum_{i=1}^{N} \left\|\boldsymbol{x}_i^j - \boldsymbol{c}^j\right\|_2, \xrightarrow{\text{Updating Rule}} \begin{array}{l} \Delta\boldsymbol{c}_t^j = \frac{\sum_{i=1}^{N} \delta\left(y_i \in j\right)\cdot\left(\boldsymbol{c}^j - \boldsymbol{x_i}\right)}{1 + \sum_{i=1}^{N}\delta\left(y_i \in j\right)}, \\ \boldsymbol{c}_{t+1}^j = \boldsymbol{c}_t^j - \alpha\cdot\Delta\boldsymbol{c}_t^j \end{array}$$

Soft Large Margin Centralized Cosine Loss

$$L_{all} = L_{LMCL} + \lambda \cdot L_{Center} + L_{Softmax} + L_{reg},$$

The overall objective loss function is a weighted sum of Softmax loss, Large Margin Cosine Loss, Regularization los and Center Loss.

To specify, center loss and softmax loss was used as the loss function for shinking the intra-class variance while the LMCL was used to enlarge the inter-class variance.

Intuitive Analysis

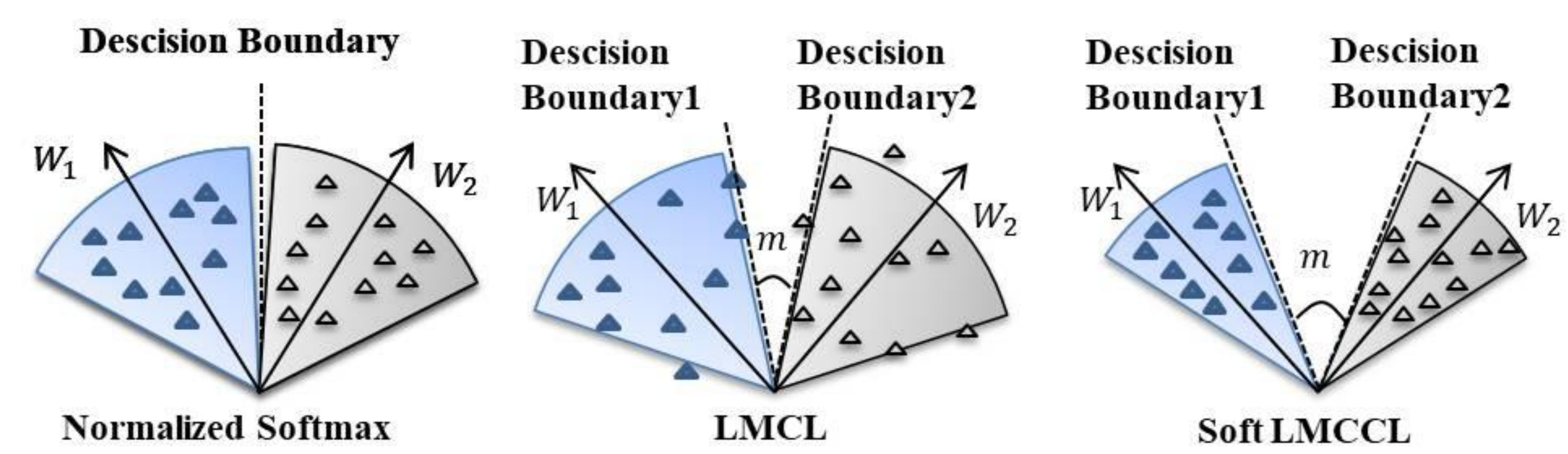


Fig. 2. Comparison of three different loss functions in binary classification.

Experiments and Results

Datasets

Simulated Datasets

1. 23 classes of subtomograms at SNR= 0.03 (S1) and SNR= 0.05 (S2) .

Experimental Datasets

1. 2,394 subtomograms from rat neuron tomogram.

Intra-domain Verification Tests

1. The unseen data are in the same domain as the training data where their imaging parameters and noise level are similar.
2. The data splitting is set to 17:6 in the simulated datasets and 3:3 in the experimental datasets.
3. Baselines: Vanilla Softmax loss, LMCL and Center Loss. Different ablation studies to empirically determine the cosine margin *m* and the weighting parameters in the loss function.

Cross-domain Verification Tests

1. 4 sets of cross-domain verification, which are S1→S2,S2→S1,S1→R1 andS2→R1.
2. Four metrics: 1) Verification accuracy 2) True Accepted Rate (TAR) under a specific False Accepted Rate (FAR) 3) Area Under Curve (AUC) 4) Equal Error Rate (EER)

Table 1: Intra-domain Subtomogram Verification on Dataset S_1

Loss Functions	Verification Acc	TAR @ (FAR=0.1%)	TAR @ (FAR=0.01%)	AUC	EER (%)
$L_{Softmax}$	0.726+-0.010	0.02514+-0.00687	0.00117+-0.00130	0.698	0.347
L_{Center}	0.667+-0.012	0.01817+-0.00548	0.00017+-0.00050	0.501	0.500
L_{LMCL}-0.35	0.742+-0.009	0.00847+-0.00240	0.00034+-0.00067	0.760	0.284
L_{LMCL}-0.5	0.772+-0.010	0.01604+-0.00481	0.00317+-0.00330	0.749	0.304
L_{LMCL}-0.65	0.742+-0.010	0.01545+-0.00599	0.00230+-0.00208	0.684	0.357
L_{LMCL}-0.7	0.753+-0.008	0.01735+-0.00476	0.00134+-0.00102	0.664	0.379
$L_{LMCL} + L_{Center}$	0.667+-0.007	0.53698+-0.00514	**0.13784+-0.01125**	0.599	0.418
$_{MCL} + 0.01 \cdot L_{Center}$	0.734+-0.010	0.51546+-0.00402	0.1050+-0.00107	0.726	0.335
$_{LMCL} + 0.1 \cdot L_{Center}$	0.743+-0.005	0.51116+-0.00595	0.09031+-0.00093	0.736	0.321
$_{LMCL} + 0.2 \cdot L_{Center}$	0.720+-0.008	0.55828+-0.00798	0.12495+-0.00478	0.670	0.376
$_{LMCL} + 0.5 \cdot L_{Center}$	0.667+-0.010	0.50046+-0.02283	0.09918+-0.00226	0.543	0.471
Soft LMCCL	**0.794+-0.008**	**0.54761+-0.01468**	0.13613+-0.01532	**0.862**	**0.222**

Table 2: Intra-domain Subtomogram Verification on Dataset S_2

Loss Functions	Verification Acc	TAR @ (FAR=0.1%)	TAR @ (FAR=0.01%)	AUC	EER(%)
$L_{Softmax}$	0.696+-0.008	0.00802+-0.00257	0.00475+-0.00125	0.667	0.380
L_{Center}	0.667+-0.012	0.03001+-0.00168	0.07157+-0.00015	0.500	0.500
L_{LMCL}-0.35	**0.746+-0.007**	0.00882+-0.00405	0.00066+-0.00110	0.734	0.323
L_{LMCL}-0.5	0.733+-0.010	0.00703+-0.00288	0.00184+-0.00162	0.681	0.362
L_{LMCL}-0.65	0.730+-0.011	0.01020+-0.00358	0.00117+-0.00168	0.660	0.384
L_{LMCL}-0.7	0.719+-0.012	0.01418+-0.00390	0.00288+-0.00191	0.645	0.385
$L_{LMCL} + L_{Center}$	0.667+-0.008	0.55115+-0.01564	0.13256+-0.01751	0.651	0.415
$_{MCL} + 0.01 \cdot L_{Center}$	0.667+-0.011	0.49875+-0.00256	0.11564+-0.02397	0.617	0.407
$_{LMCL} + 0.1 \cdot L_{Center}$	0.688+-0.006	0.51403+-0.00340	0.10083+-0.00083	0.633	0.402
$_{LMCL} + 0.2 \cdot L_{Center}$	0.667+-0.005	0.55997+-0.00553	0.10000+-0.00672	0.545	0.475
$_{LMCL} + 0.5 \cdot L_{Center}$	0.682+-0.007	0.50218+-0.00151	0.08756+-0.00124	0.638	0.389

Table 4: Cross-domain Subtomogram Verification Results

Loss Functions	Verification Acc	TAR @ (FAR=0.1%)	TAR @ (FAR=0.01%)	AUC	EER(%)
		$S_1 \to S_2$			
$L_{Softmax}$	0.730+-0.011	0.01867+-0.00254	0.00564+-0.00025	0.728	0.329
L_{Center}	0.667+-0.011	0.03265+-0.01254	0.01002+-0.00000	0.500	0.500
L_{LMCL}-0.5	**0.789+-0.009**	0.02521+-0.00542	0.00533+-0.00336	0.799	0.263
Soft LMCCL	0.788+-0.005	**0.50930+-0.01185**	**0.11786+-0.00790**	**0.862**	**0.225**
		$S_2 \to S_1$			
$L_{Softmax}$	0.694+-0.009	0.01710+-0.00595	0.00256+-0.00012	0.659	0.381
L_{Center}	0.667+-0.011	0.05135+-0.00145	0.00425+-0.00151	0.500	0.500
L_{LMCL}-0.35	0.731+-0.009	0.00833+-0.00404	0.00083+-0.00083	0.697	0.343
Soft LMCCL	**0.734+-0.007**	**0.53915+-0.00732**	**0.10301+-0.00238**	**0.784**	**0.293**
		$S_1 \to R_1$			
$L_{Softmax}$	0.554+-0.012	0.03428+-0.00409	0.00672+-0.00138	0.503	0.489
L_{Center}	0.554+-0.019	0.01439+-0.00147	0.01125+-0.00243	0.552	0.462
L_{LMCL}-0.5	0.619+-0.017	0.01708+-0.00607	0.00838+-0.00399	0.525	0.489
Soft LMCCL	**0.637+-0.011**	**0.50668+-0.00263**	**0.09066+-0.00133**	**0.663**	**0.371**
		$S_2 \to R_1$			
$L_{Softmax}$	0.565+-0.021	0.11230+-0.00514	0.07519+-0.02561	0.573	0.436
L_{Center}	0.554+-0.015	0.01025+-0.00002	0.00108+-0.00727	0.500	0.500
L_{LMCL}-0.35	**0.647+-0.020**	0.09845+-0.00771	0.01564+-0.00357	0.658	0.378
Soft LMCCL	0.637+-0.017	**0.52199+-0.01184**	**0.11635+-0.00838**	**0.669**	**0.373**

Fig. 3. Experimental results on simulated datasets and real datasets for intra-domain verification tests and cross-domain verification tests