Appendix

Author Name
Affiliation

1 Riemannian Stochastic Gradient Descent

Riemannian Stochastic Gradient Descent (RSGD) [Bonnabel, 2013] is a technique that updates parameters by back-propagation in hyperbolic spaces. Here we briefly introduce the RSGD process applied by [Nickel and Kiela, 2017] in a Poincaré ball.

Suppose we are given a set $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i\}_{i=1}^n$ of parameters in a Poincaré ball of radius 1: $\forall \boldsymbol{\theta}_i \in \boldsymbol{\Theta}$, $\|\boldsymbol{\theta}_i\| < 1$. Given a loss function $L(\boldsymbol{\Theta})$, We aim at optimizing $\boldsymbol{\Theta}$ by minimizing $L(\boldsymbol{\Theta})$:

$$\boldsymbol{\Theta}^{'} \leftarrow \operatorname*{argmin}_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta})$$

We should first calculate the Riemannian gradients for every parameters, and then update the parameters in the Poincaré ball. given the learning rate η, the back propagation in the Poincaré ball can be defined as:

$$\boldsymbol{\theta}_{t+1} \leftarrow P_{\boldsymbol{\theta}_t}(-\eta \nabla_R L(\boldsymbol{\theta}_t)),$$

where P is a retraction function, ∇_R is the Riemannian gradient. The retraction is related to both $\eta \nabla_R L(\boldsymbol{\theta}_t)$ and $\boldsymbol{\theta}$'s position at time t, and its function is similar to the optimization step in Euclidean space simply achieved by:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \nabla_E L(\boldsymbol{\theta}_t).$$

The Riemannian gradient is not difficult to calculate. Given the Riemannian metric tensor: $g_{\boldsymbol{\theta}_t} = \lambda_{\boldsymbol{\theta}_t}^2 I$ where $\lambda_{\boldsymbol{\theta}_t} = \frac{2}{1-\|\boldsymbol{\theta}_t\|^2}$ and I being the identity matrix, the Riemannian gradient can be calculated based on the Euclidean gradient ∇_E:

$$\nabla_R L(\boldsymbol{\theta}_t) = g_{\boldsymbol{\theta}_t}^{-1} \nabla_E L(\boldsymbol{\theta}_t).$$

Therefore, given the loss function L, we can first calculate the Euclidean gradient ∇_E of the parameter $\boldsymbol{\theta}$ at time t by traditional back-propagation in Euclidean space, and then divided it by $g_{\boldsymbol{\theta}_t}$ to obtain the Riemannian gradient of $\boldsymbol{\theta}_t$.

Given the Riemannian gradient $\nabla_R L(\boldsymbol{\theta}_t)$, Nickel and Kiela define their updating process as:

$$\boldsymbol{\theta}_{t+1} \leftarrow Q(\boldsymbol{\theta}_t - \eta \nabla_R L(\boldsymbol{\theta}_t)),$$

where

$$Q(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} / \|\boldsymbol{x}\| - \epsilon & \|\boldsymbol{x}\| \geq 1 \\ \boldsymbol{x} & \|\boldsymbol{x}\| < 1 \end{cases},$$

and ϵ serve as a small vector. In this way, if $\boldsymbol{\theta}_{t+1}$ falls out of the Poincaré ball with radius 1, ϵ will pull it back into the ball, therefore $\boldsymbol{\theta}_t$ always satisfies $\|\boldsymbol{\theta}_t\| < 1$.

The drawback of this linear retraction is that it neglects the characteristics of the hyperbolic spaces. By contrast, We use exponential map to perform the retraction. Exponential and logarithmic maps serve as transformation tools between a Poincaré Ball $\mathbb{D}^{d,1}$ and an Euclidean tangent space $\mathbb{E}^d$. For any point $\boldsymbol{x} \in \mathbb{D}^{d,1}$, the exponential map and the logarithm map for $\boldsymbol{v} \neq \mathbf{0}$ and $\boldsymbol{y} \neq \boldsymbol{x}$ are:

$$\exp_{\boldsymbol{x}}(\boldsymbol{v}) = \boldsymbol{x} \oplus (\tanh(\frac{\lambda_{\boldsymbol{x}} \|\boldsymbol{v}\|}{2}) \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}),$$

$$\log_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{2}{\lambda_{\boldsymbol{x}}} \tanh^{-1}(\|-\boldsymbol{x} \oplus \boldsymbol{y}\|) \frac{-\boldsymbol{x} \oplus \boldsymbol{y}}{\|-\boldsymbol{x} \oplus \boldsymbol{y}\|},$$

where

$$\boldsymbol{x} \oplus \boldsymbol{y} = \frac{(1 + 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \|\boldsymbol{y}\|^2)\boldsymbol{x} + (1 - \|\boldsymbol{x}\|^2)\boldsymbol{y}}{1 + 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2}.$$

The Riemannian gradient $\nabla_R L(\boldsymbol{\theta}_t)$ can be viewed as a vector in the tangent space of $\boldsymbol{\theta}_t$. Therefore, we can perform the optimization step on $\boldsymbol{\theta}_t$ by mapping the Riemannian gradient from $\boldsymbol{\theta}_t$'s tangent space to the Poincaré ball to obtain $\boldsymbol{\theta}_{t+1}$. Formally, we get $\boldsymbol{\theta}_{t+1}$ by:

$$\boldsymbol{\theta}_{t+1} \leftarrow \exp_{\boldsymbol{\theta}_t}(-\nabla_R L(\boldsymbol{\theta}_t)).$$

2 Experiments

2.1 Dataset Construction

Here we briefly describe how we construct our datasets.

Protein-GO. We use the protein-protein interaction network (PPI) of humans provided by STRING [Szklarczyk *et al.*, 2021]. We set a threshold of 800 to sample the edges in the original data to form our PPI network. We then delete the proteins not appearing in GOA [Camon *et al.*, 2004; Huntley *et al.*, 2015], so that each protein has its GO-terms in the GO taxonomy.

Gene-Pathway. We use the human gene regulatory network provided by [Liu *et al.*, 2015] and the pathway ontology provided by [Petri *et al.*, 2014]. We pick the pathways that have Reactome ID from the original pathway ontology

to form our taxonomy of 241 labels. We then use the gene-pathway associations from CTD [Davis *et al.*, 2020] to connect pathways to genes.

DBLP-ACM. We sample a subset of the DBLP network with the research taxonomy constructed by [Yang *et al.*, 2020], where each author has at least one co-author in the network. Every author in the network has at least one label in the research taxonomy.

2.2 Parameter Settings

On all datasets, we set the learning rate of the residual vectors $\boldsymbol{c}$ and the label vectors $\mathcal{Q}$ to 0.01, while for the branch vectors B we set their learning rate to 0.02. We set the learning rate of p to 0.1. The negative sampling number is set to 5, and the batch size is 1000. The LRU period is set to 5 epochs. We train HIME on all datasets for 100 epochs.

2.3 Visualization Results

We provide the visualization results of HIME_4 on the three datasets, as shown in figure 1,2, 3 and 4.

References

[Bonnabel, 2013] Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[Camon *et al.*, 2004] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic acids research*, 32(suppl_1):D262–D266, 2004.

[Davis *et al.*, 2020] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143, 10 2020.

[Huntley *et al.*, 2015] Rachael P Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J Martin, and Claire O'Donovan. The goa database: gene ontology annotation updates for 2015. *Nucleic acids research*, 43(D1):D1057–D1063, 2015.

[Liu *et al.*, 2015] Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015, 09 2015. bav095.

[Nickel and Kiela, 2017] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

[Petri *et al.*, 2014] Victoria Petri, Pushkala Jayaraman, Marek Tutaj, G. Thomas Hayman, Jennifer R. Smith, Jeff De Pons, Stanley JF Laulederkind, Timothy F. Lowry, Rajni Nigam, Shur-Jen Wang, Mary Shimoyama, Melinda R. Dwinell, Diane H. Munzenmaier, Elizabeth A. Worthey, and Howard J. Jacob. The pathway ontology – updates and applications. *Journal of Biomedical Semantics*, 5(1):7, Feb 2014.

[Szklarczyk *et al.*, 2021] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[Yang *et al.*, 2020] Carl Yang, Jieyu Zhang, and Jiawei Han. Co-embedding network nodes and hierarchical labels with taxonomy based generative adversarial networks. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 721–730. IEEE, 2020.

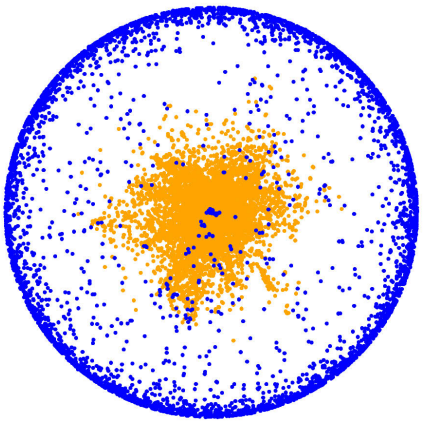

Figure 1: The 2D embeddings of HIME_4 on Protein-GO. Blue points are labels and the orange points are the active branch vectors of all nodes.

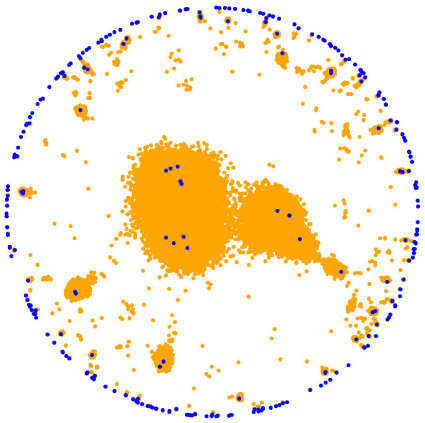

Figure 3: The 2D embeddings of HIME_4 and EUHIME_4 on DBLP-ACM. Blue points are labels and the orange points are the active branch vectors of all nodes.

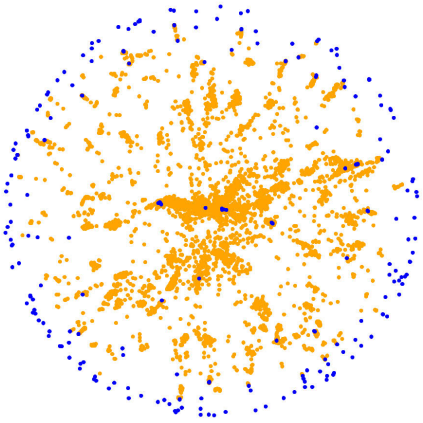

Figure 2: The 2D embeddings of HIME_4 and EUHIME_4 on Gene-Pathway. Blue points are labels and the orange points are the active branch vectors of all nodes.

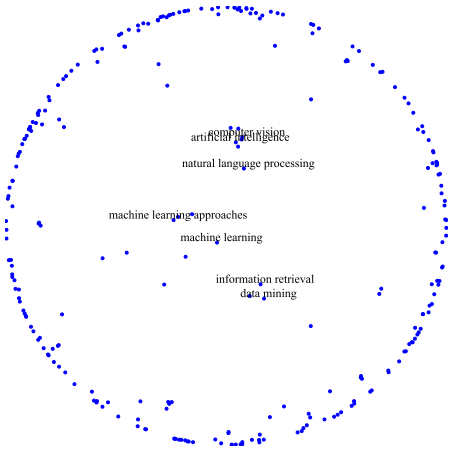

Figure 4: The 2D hierarchical label embeddings of HIME_4 on DBLP-ACM.