

Supplementary Materials: DySarL: Dynamic Structure-Aware Representation Learning for Multimodal Knowledge Graph Reasoning

Anonymous Authors

1 HYPERBOLIC GEOMETRY

Hyperbolic geometry [1, 7] is a non-Euclidean geometry with negative curvature, while Euclidean geometry has zero curvature. Curvature measures the extent to which a point deviates from a plane. Due to the superlinear growth of distances on hyperbolic manifolds, hyperbolic geometry can effectively express the hierarchical structures within low-dimensional limited Euclidean neighborhoods [4, 5].

Inspired by hyperbolic embedding-based approaches [3, 4], the design of hyperbolic messages that perceive hierarchy-based multihop structural features over MKGs in the dual-space multihop structural learning (DMS) module of our proposed DySarL model extensively utilizes hyperbolic geometry. Hence, we present some basic formulas to facilitate a thorough understanding of Equations 2, 3, and 11 in Section 3.

In Equation 2, $\exp_{c_r}(\cdot)$ and $\log_{c_r}(\cdot)$ denote the exponential and logarithmic mapping operations with reference to the hyperbolic origin $\mathbf{0}$, which project the entity embeddings from the Euclidean space (\mathbb{R}) to the hyperbolic space (\mathbb{B}) and from the hyperbolic space (\mathbb{B}) back to the Euclidean space (\mathbb{R}), respectively. They can be formulated as follows:

$$\exp_{c_r}(\mathbf{u}) = \tanh(\sqrt{c_r}\|\mathbf{u}\|) \frac{\mathbf{u}}{\sqrt{c_r}\|\mathbf{u}\|} \quad (1)$$

$$\log_{c_r}(\mathbf{u}^{\mathbb{B}}) = \operatorname{artanh}(\sqrt{c_r}\|\mathbf{u}^{\mathbb{B}}\|) \frac{\mathbf{u}^{\mathbb{B}}}{\sqrt{c_r}\|\mathbf{u}^{\mathbb{B}}\|} \quad (2)$$

where $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{u}^{\mathbb{B}} \in \mathbb{B}^d$ refer to specific embeddings (points) in the Euclidean space and hyperbolic space, respectively. c_r denotes the learnable relation-specific curvatures. $\|\cdot\|$ denotes the L2 normalization operation. \oplus^{c_r} indicates the Möbius addition operation [6], which can be formulated as follows:

$$\mathbf{u}^{\mathbb{B}} \oplus^{c_r} \mathbf{v}^{\mathbb{B}} = \frac{\left(1 + 2c_r\langle \mathbf{u}^{\mathbb{B}}, \mathbf{v}^{\mathbb{B}} \rangle + c_r\|\mathbf{v}^{\mathbb{B}}\|^2\right) \mathbf{u}^{\mathbb{B}} + \left(1 - c_r\|\mathbf{u}^{\mathbb{B}}\|^2\right) \mathbf{v}^{\mathbb{B}}}{1 + 2c_r\langle \mathbf{u}^{\mathbb{B}}, \mathbf{v}^{\mathbb{B}} \rangle + c_r^2\|\mathbf{u}^{\mathbb{B}}\|^2\|\mathbf{v}^{\mathbb{B}}\|^2} \quad (3)$$

where $\mathbf{u}^{\mathbb{B}}$ and $\mathbf{v}^{\mathbb{B}} \in \mathbb{B}^d$ represent specific embeddings (points) in the hyperbolic space. $\langle \cdot \rangle$ denotes the dot product operation.

Then, in Equation 3, $\operatorname{Rot}(\Theta_r)$ and $\operatorname{Ref}(\Phi_r)$ are block diagonal matrices that represent the rotation and reflection operations, respectively. Θ_r and Φ_r are relation-specific parameters. Specifically, $\operatorname{Rot}(\Theta_r) = \operatorname{diag}(G^+(\Theta_{r,1}), \dots, G^+(\Theta_{r,\frac{d}{2}}))$, and $\operatorname{Ref}(\Phi_r) = \operatorname{diag}(G^-(\Phi_{r,1}), \dots, G^-(\Phi_{r,\frac{d}{2}}))$, where $G^\pm(\cdot)$ denotes the given transformations [3] in the form of 2×2 matrices:

$$G^+(\Theta_{r,i}) = \begin{bmatrix} \cos \Theta_{r,i} & -\sin \Theta_{r,i} \\ \sin \Theta_{r,i} & \cos \Theta_{r,i} \end{bmatrix} \quad (4)$$

$$G^-(\Phi_{r,i}) = \begin{bmatrix} \cos \Phi_{r,i} & \sin \Phi_{r,i} \\ \sin \Phi_{r,i} & -\cos \Phi_{r,i} \end{bmatrix} \quad (5)$$

$\operatorname{Att}(\cdot)$ in Equation 3 learns an appropriate combination of hyperbolic reflections and rotations through a learnable attention weight \mathbf{a}_r . This process can be formulated as follows:

$$\operatorname{Att}(\mathbf{u}^{\mathbb{B}}, \mathbf{v}^{\mathbb{B}}; \mathbf{a}_r) = \exp_{c_r}(f(\mathbf{a}_r^T \mathbf{u})\mathbf{u} + f(\mathbf{a}_r^T \mathbf{v})\mathbf{v}) \quad (6)$$

where $\mathbf{u} = \log_{c_r}(\mathbf{u}^{\mathbb{B}})$ and $\mathbf{v} = \log_{c_r}(\mathbf{v}^{\mathbb{B}})$. $f(\cdot)$ is the Softmax activation function.

Finally, in Equation 11, for a certain multimodal query $(s, r, ?)$, ATTH [3] calculates the hyperbolic distances between all candidate entities and the Möbius summation of s and r , assigning higher scores to entities that are closer in distance. This process can be formulated as follows:

$$\mathcal{S} = -d^{c_r}((H_r(s^{\mathbb{B}}) \oplus^{c_r} r^{\mathbb{B}}), \mathbf{H})^2 + b_s + b_o \quad (7)$$

where $H_r(s^{\mathbb{B}}) \in \mathbb{B}^d$ refers to Equation 3. $s^{\mathbb{B}}$ and $r^{\mathbb{B}} \in \mathbb{B}^d$ are hyperbolic embeddings of entity s and relation r . $\mathbf{H} \in \mathbb{B}^{N \times d}$ is the embedding matrix in the hyperbolic space of all entities corresponding to the Euclidean embeddings in \mathbf{E} (or \mathbf{E}_U). $b_s \in \mathbb{R}$ and $b_o \in \mathbb{R}^N$ are the learnable biases for subject s and candidate objects o , respectively. Moreover, the hyperbolic distance $d^{c_r}(\cdot)$ is formulated as follows:

$$d^{c_r}(\mathbf{u}^{\mathbb{B}}, \mathbf{v}^{\mathbb{B}}) = \frac{2}{\sqrt{c_r}} \operatorname{artanh}\left(\sqrt{c_r}\left\|-\mathbf{u}^{\mathbb{B}} \oplus^{c_r} \mathbf{v}^{\mathbb{B}}\right\|\right) \quad (8)$$

where $\mathbf{u}^{\mathbb{B}}$ and $\mathbf{v}^{\mathbb{B}} \in \mathbb{B}^d$.

2 CROSSMODAL ATTENTIVE DYNAMICS

Cross-modal attentive dynamics in MKG reasoning refer to the dynamic changes in the attentional emphasis of different reasoning facts on different modal features. It signifies the dynamic role played by different modal features in predicting different reasoning facts. Here, we further explain the two aspects involved to facilitate the understanding of cross-modal attentive dynamics.

Different modal features play dynamic (distinct) roles in reasoning for a specific decision fact. For example, for the reasoning fact (*Joe Biden, LiveAt, ?*), the structural neighborhood features of the entity “Joe Biden” and the visual and linguistic features derived from his pictures and resumes respectively play different roles in the prediction process. As mentioned in Section 1, previous coattention-based cross-modal fusion methods [2, 8] in MKG reasoning have primarily focused on the interplay between different modalities, making it challenging to effectively capture the dynamic effects of different modal features on reasoning facts. In our proposed interactive symmetric attention fusion (ISA) module, the symmetric attention component is carefully designed to introduce an initialized feature matrix \mathbf{E}_U as the attention sender. It then symmetrically and uniformly treats all modal feature matrices $\{\mathbf{E}_S, \mathbf{E}_I, \mathbf{E}_T\}$ in MKGs as attention targets for factual inference

(see Equation 9). Hence, DySarI effectively addresses the 1st aspect of the cross-modal attentive dynamics issue in MKG reasoning.

A specific type of modal features plays dynamic (distinct) roles in reasoning for different decision facts. This phenomenon is more pronounced in predicting different facts that involve the same entity. For instance, when predicting (*Joe Biden, LiveAt, ?*) and (*Joe Biden, Visit, ?*), the structural modal features that need to be aggregated from the neighborhoods of the “*Joe Biden*” entity should include more information related to residential buildings and political countries to highlight the ground-truth entities “*White House*” and “*Ukraine*” respectively in the final prediction. Thus, for the same entity “*Joe Biden*” in different factual predictions such as (*Joe Biden, LiveAt, ?*) and (*Joe Biden, Visit, ?*), a specific type of modal features (e.g., the structural modality) exerts different influences. As illustrated in Figure 1(b), traditional codec-based architectures fail to effectively address this problem because they rely on fixed, static multimodal embeddings learned by the encoder for score calculation during the decoding process. In our proposed ISA module, the carefully designed fact-specific gated attention unit establishes a connection between the cross-modal feature fusion process in the encoding stage and the score calculation process in the decoding stage by introducing two learnable parameters $\{\delta_s, \delta_r\}$ during the inference (decoding) phase. Specifically, the δ_s and δ_r parameters provide the capability to finetune the scores derived from the attention sender and attention targets based on specific entities (e.g., “*Joe Biden*”) and relations (e.g., “*LiveAt*” and “*Visit*”) for the inference facts. Therefore, DySarI effectively addresses the 2nd aspect of the cross-modal attentive dynamics issue in MKG reasoning.

In the experiments conducted in Section 4.5, we demonstrate that our proposed DySarI model takes into account both the above-mentioned aspects, thereby effectively capturing the cross-modal attentive dynamics in MKG reasoning.

REFERENCES

- [1] Ben Andrews and Christopher Hopper. 2010. *The Ricci flow in Riemannian geometry: a complete proof of the differentiable 1/4-pinching sphere theorem*. Springer.
- [2] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- [3] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. 2020. Low-Dimensional Hyperbolic Knowledge Graph Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. ACL, 6901–6914.
- [4] Yan Jia, Mengqi Lin, Ye Wang, Jianming Li, Kai Chen, Joanna Siebert, Geordie Z. Zhang, and Qing Liao. 2023. Extrapolation over temporal knowledge graph via hyperbolic embedding. *CAAI Trans. Intell. Technol.* 8, 2 (2023), 418–429.
- [5] Qi Liu, Maximilian Nickel, and Douwe Kiela. 2019. Hyperbolic Graph Neural Networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.), 8228–8239. <https://proceedings.neurips.cc/paper/2019/hash/103303dd56a731e377d01f6a37badae3-Abstract.html>
- [6] Abraham A Ungar. 2001. Hyperbolic trigonometry and its application in the Poincaré ball model of hyperbolic geometry. *Computers & Mathematics with Applications* 41, 1-2 (2001), 135–147.
- [7] Thomas James Willmore. 2013. *An introduction to differential geometry*. Courier Corporation.
- [8] Shangfei Zheng, Weiqing Wang, Jianfeng Qu, Hongzhi Yin, Wei Chen, and Lei Zhao. 2023. MMKGR: Multi-hop Multi-modal Knowledge Graph Reasoning. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 96–109.

175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232