Language Models as Hierarchy Encoders

Yuan He[1], Zhangdie Yuan[2], Jiaoyan Chen[3,1], Ian Horrocks[1]

[1]University of Oxford, [2]University of Cambridge, [3]University of Manchester

Motivation

Existing pre-trained LMs lack explicit hierarchy interpretation

- Pre-trained LMs can predict relations like "A is B" and "B is C" but struggle to infer the **transitive relationship** "A is C" [1]
- These models typically encode hierarchical entities based on **similarities** rather than structural relationships [2]

Limitations of existing hyperbolic embeddings

- Classic hyperbolic embeddings, such as Poincare Embeddings [3] and Hyperbolic Entailment Cone [4], are **static** and only capture hierarchy within a **fixed entity set**
- Hyperbolic word embeddings [5] face limitations due to **word-level tokenisation** and **unified word representations**

Preliminaries

Hyperbolic Geometry

- A form of **non-Euclidean geometry** characterised by its constant negative Gaussian curvature
- The distance between points grows **exponentially** as they approach the manifold's boundary
- Provides a **theoretical guarantee** for embedding tree-like structures [4]
- **Poincaré ball**: A d-dimensional open ball $\mathbb{B}_c^d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|^2 < \frac{1}{c}\}$
- **Distance function**: $d_c(\mathbf{u}, \mathbf{v}) = \frac{2}{\sqrt{c}} \tanh^{-1}(\sqrt{c}\|- \mathbf{u} \oplus_c \mathbf{v}\|)$ where $\oplus_c$ denotes the Möbius addition.

Hierarchy

- A directed acyclic graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ represents **entities** as vertices and $\mathcal{E}$ represents **direct subsumption relationships** as edges
- **Indirect subsumptions** $\mathcal{T}$ are derived from *transitive reasoning*
- **Negative subsumptions** are $(e_1 \in \mathcal{E}, e_2 \in \mathcal{E}) \notin \mathcal{E} \cup \mathcal{T}$ (closed-world assumption)

References

[1] Lin et al. "Does bert know that the is-a relation is transitive?" In: ACL 2022.

[2] Liu et al. "Self-alignment pretraining for biomedical entity representations" In: NAACL 2021.

[3] Nickel et al. "Poincaré embeddings for learning hierarchical representations" In: NeurIPS 2017.

[4] Ganea et al. "Hyperbolic entailment cones for learning hierarchical embeddings" In: ICML 2018.

[5] Tifrea et al. "Poincare glove: Hyperbolic word embeddings." In: ICLR 2018.

Huggingface Hub

Hierarchy Transformer Encoder (HIT)

Construction

- The output embedding space of transformer encoder-based LMs is often a ***d*-dimensional hyper-cube** due to the $\tanh$ activation function in the last layer. We can then construct a **Poincaré ball** of radius $\sqrt{d}$ (a **d-dimensional hyper-sphere**) so that its boundary **circumscribes** the output embedding space of LMs
- We utilise the **sentence transformer architecture** except that we **exclude the normalisation layer** after mean pooling over token embeddings as it prevents hierarchical organisation

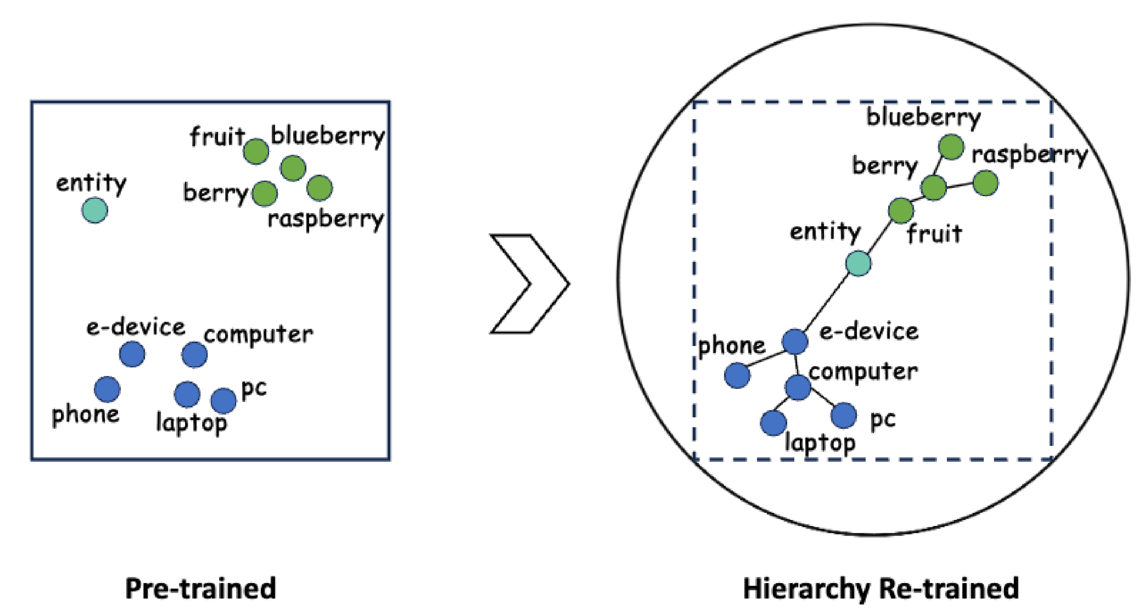


Fig 1. Illustration of how hierarchies are explicitly encoded in **HIT**.

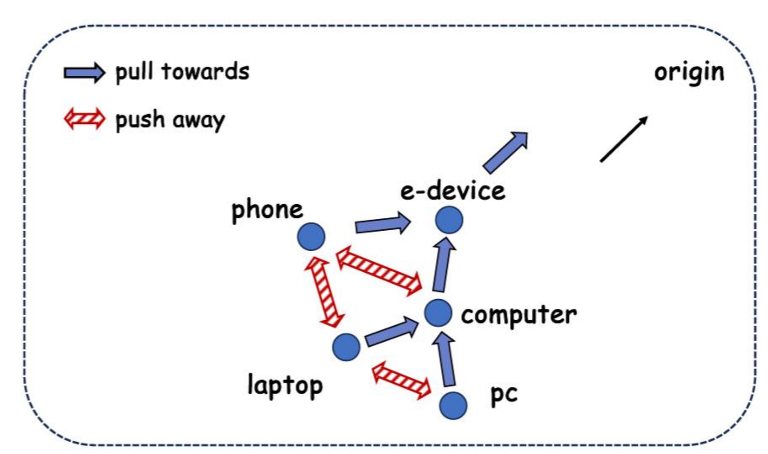


Fig 2. Illustration of the impact of hyperbolic clustering and centripetal losses.

Hyperbolic Losses

- **Hyperbolic Clustering Loss**: to cluster related entities while distancing unrelated ones

$$\mathcal{L}_{cluster} = \sum_{(e, e^+, e^-) \in \mathcal{D}} \max(d_c(\mathbf{e}, \mathbf{e}^+) - d_c(\mathbf{e}, \mathbf{e}^-) + \alpha, 0)$$

- **Hyperbolic Centripetal Loss**: to position the parent entities closer to the manifold's origin than child counterparts

$$\mathcal{L}_{centri} = \sum_{(e, e^+, e^-) \in \mathcal{D}} \max(\|\mathbf{e}^+\| - \|\mathbf{e}\| + \beta, 0)$$

- The overall loss is the linear combination of the above two losses.
- **Subsumption Prediction Function**: to probe the resulting **HIT** model to predict entity subsumptions

$$s(e_1 \sqsubseteq e_2) = -(d_c(\mathbf{e}_1, \mathbf{e}_2) + \lambda(\|\mathbf{e}_2\|_c - \|\mathbf{e}_1\|_c))$$

Evaluation

Task Definition

- **Multi-hop Inference**: Trained on asserted (one-hop) subsumptions and tested on transitively inferred (multi-hop) subsumptions
- **Mixed-hop Prediction**: Trained on incomplete asserted subsumptions and tested on arbitrary, probably unseen subsumptions
- **Mixed-hop Prediction (Transfer)**: Trained on asserted subsumptions of one hierarchy and tested on arbitrary subsumptions of another hierarchy
- **Evaluation Metrics**: Precision, Recall, and F-score

Dataset

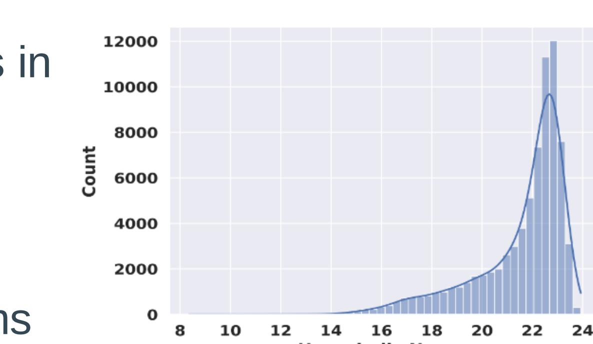
Source	#Entity	#DirectSub	#IndirectSub	#Dataset (Train/Val/Test)
WordNet	74,401	75,850	587,658	multi: 834K/323K/323K mixed: 751K/365K/365K
Schema.org	903	950	1,978	mixed: -/15K/15K
FoodOn	30,963	36,486	438,266	mixed: 361K/261K/261K
DOID	11,157	11,180	45,383	mixed: 122K/31K/31K
SNOMED	364,352	420,193	2,775,696	mixed: 4,160K/1,758K/1,758K

Results

	Random Negatives			Hard Negatives		
Model	Precision	Recall	F-score	Precision	Recall	F-score
NaivePrior	0.091	0.091	0.091	0.091	0.091	0.091
Multi-hop Inference (WordNet)						
PoincaréEmbed	0.862	0.866	0.864	0.797	0.867	0.830
HyperbolicCone	0.817	0.996	0.898	0.243	0.902	0.383
all-MiniLM-L12-v2	0.127	0.585	0.209	0.108	0.740	0.188
+ fine-tune	0.811	0.515	0.630	0.819	0.530	0.643
+ HIT	0.880	0.927	0.903	0.910	0.906	0.908
Mixed-hop Prediction (WordNet)						
all-MiniLM-L12-v2	0.127	0.583	0.209	0.111	0.625	0.188
+ fine-tune	0.794	0.517	0.627	0.859	0.515	0.644
+ HIT	0.875	0.895	0.885	0.886	0.857	0.871
Transfer Mixed-hop Prediction (WordNet → DOID)						
PoincaréGloVe	0.265	0.314	0.287	0.283	0.318	0.299
all-MiniLM-L12-v2	0.342	0.451	0.389	0.159	0.455	0.235
+ fine-tune	0.585	0.621	0.603	0.868	0.179	0.297
+ HIT	0.696	0.711	0.704	0.810	0.435	0.566

Analysis

- The hyperbolic norms of entity embeddings in **HIT** capture the natural expansion of hierarchies
- **HIT** demonstrates a stronger linear relationship between entity hyperbolic norms and depths



HIT	PoincaréEmbed	HyperbolicCone
0.346	0.130	0.245

Future Work

- Mitigate catastrophic forgetting
- Develop **hierarchy-based** semantic search