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

- Do current pre-trained language models (LMs) encode **hierarchical information** explicitly and effectively?
- **Not explicitly**. Many LMs are optimised on **text similarity** for semantic search and paraphrasing [Reimers et al. EMNLP'19, Liu et al. NAACL'21]
- **Not effectively**. LMs fail to capture **transitivity** of "is-a" [lin et al. ACL'22]

Our Contributions

- An approach to **re-train** (diff. from standard fine-tuning) **transformer encoder-based LMs** as explicit **Hierarchy Transformer encoders (HiTs)**, utilising the expansive nature of hyperbolic geometry

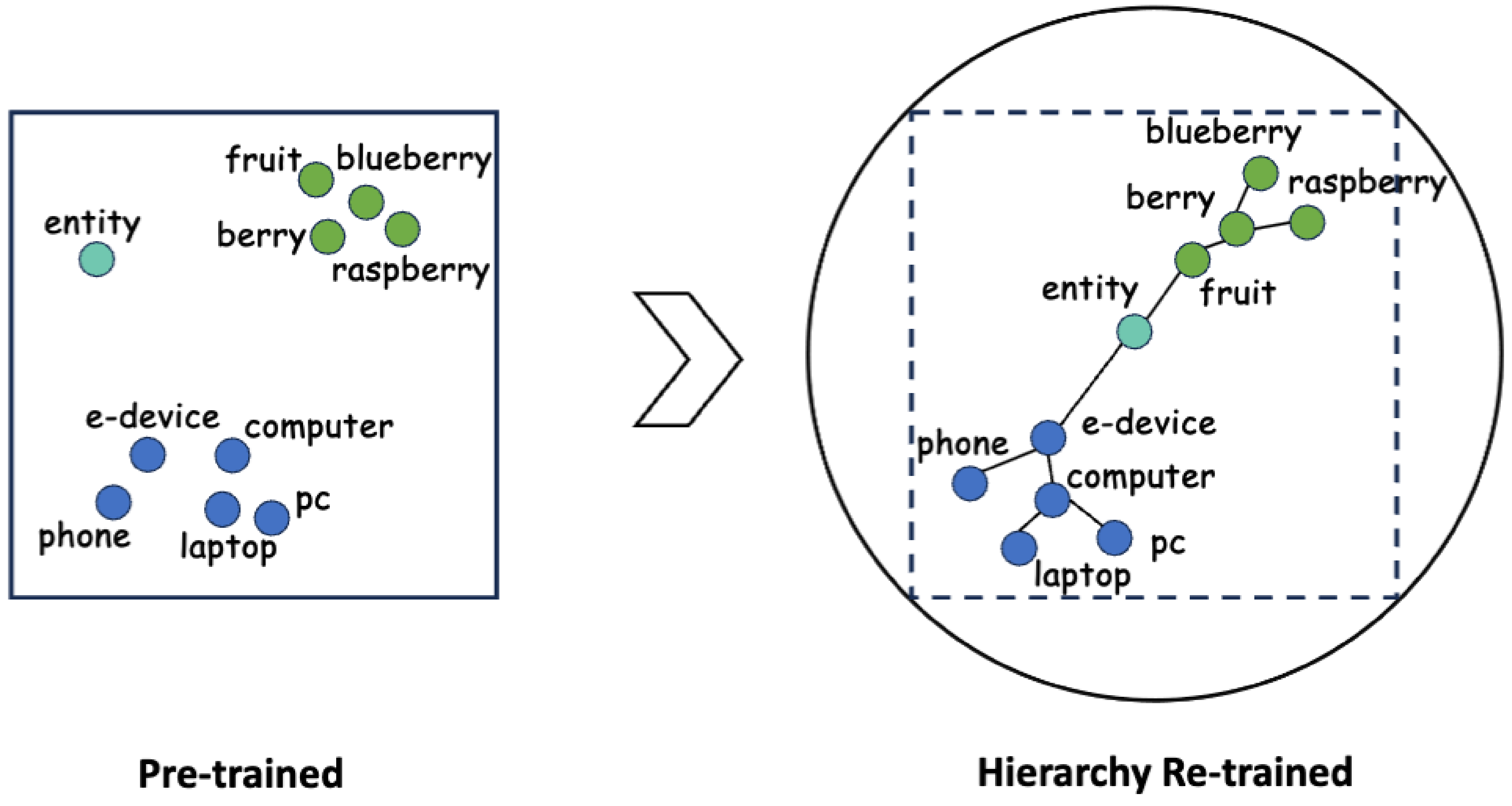
- Our results show:
  - More effective than pre-trained and (standard) fine-tuned LMs
  - More effective than previous hyperbolic embedding methods and support **inductive predictions** within and across hierarchies
  - Embeddings demonstrate **geometric interpretability**

Hyperbolic Geometry

- **Constant negative curvature** (as opposed to flat, zero curvature in Euclidean geometry)
- Usually defined by an open set with a metric tensor conformal (same angle measurement) to Euclidean geometry.
  - E.g., **Poincaré ball** is defined by an open ball $\mathrm{B}_c^d = \{x \in \mathbb{R}^d : \| x \| < \frac{1}{c}\}$ where c is the curvature value
- **Distances grow exponentially** as approaching towards the boundary → naturally follows the expansion of hierarchy
- **Theoretical property** for embedding tree-like structure: δ-hyperbolicity

Methodology

- **Square**: LMs' last output activation is $\tanh$, mapping each dimension to $[-1, 1]$.
- **Circle**: Poincaré ball of a negative curvature $-1/d$ that circumscribes LMs' output embedding space.



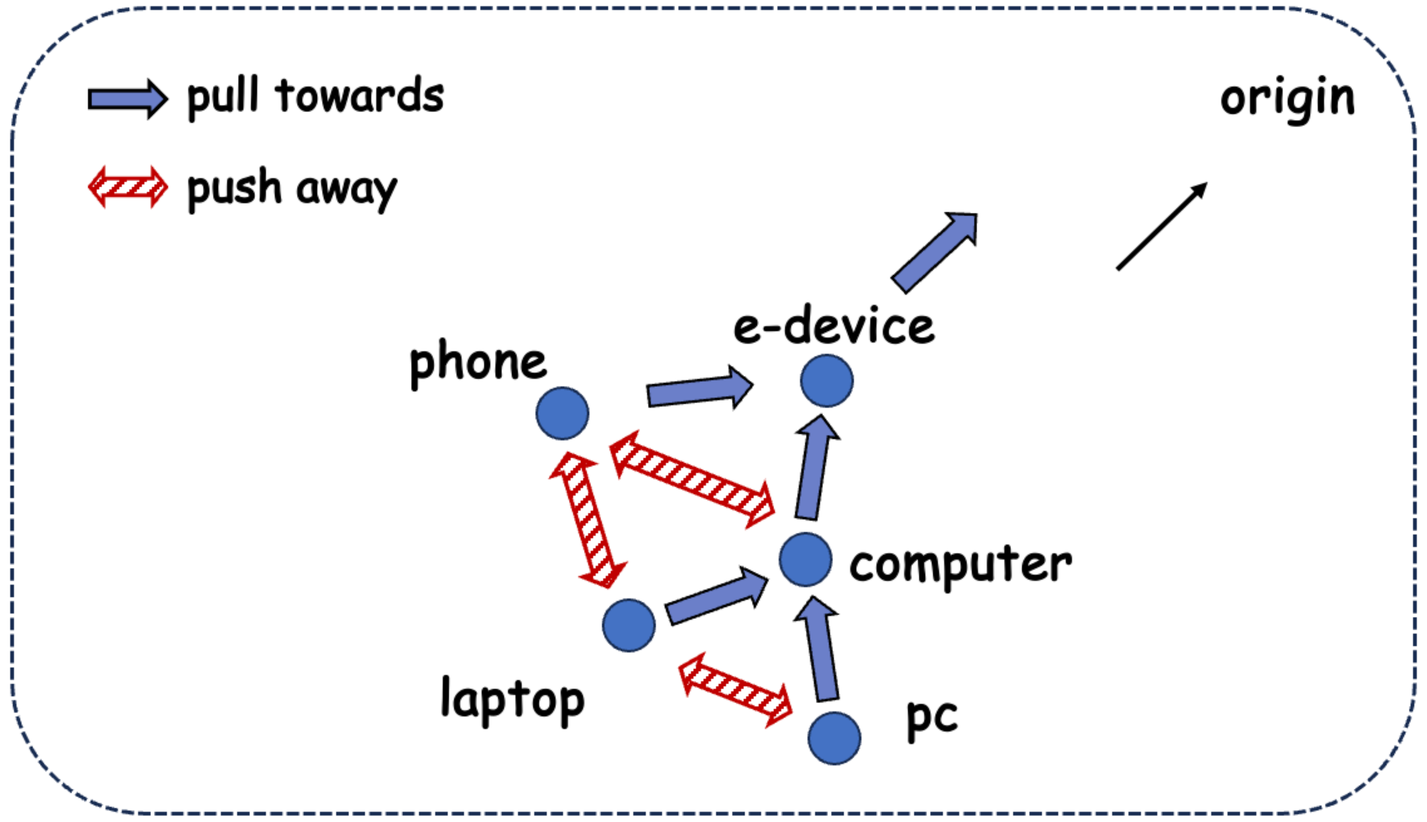
Methodology

- **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.

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



Methodology

- **Overall loss** is the linear combination of the two hyperbolic losses.

- **Subsumption Prediction Function**: probe **HIT** models to predict entity subsumptions

$$s(e_1 \sqsubseteq e_2) = -(d_c(e_1, e_2) + \lambda\,(\| e_2 \|_c - \| e_1 \|_c))$$

where c is curvature, λ is the weight for hyperbolic norm diff.

Task Definitions

Multi-hop Inference: We define base edges as those asserted in the hierarchy; the task aims to infer *transitive (multi-hop)* edges from base edges.

Mixed-hop Prediction: We split base edges for training and testing. The test set represents missing subsumptions (which may lead to unseen entities). Models are required to predict subsumptions between *arbitrary (mixed-hop)* entity pairs.

Mixed-hop Prediction (Transfer): Trained on base edges of one hierarchy and tested on arbitrary entity pairs of another.

Datasets

- **Main Hierarchies:** WordNet, SNOMED-CT
- **Transfer Evaluation:** Schema.org, FoodOn, DOID

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

Baselines

Pre-trained LMs: since no LMs are trained on encoding hierarchies explicitly, we design probes that follow the pre-training objectives for prediction.

Fine-tuned LMs: attaching a downstream layer for end-to-end classification.

Hyperbolic Baselines: Poincaré Embedding [Nickel et al. NeurIPS'17], Hyperbolic Entailment Cone [Ganea et al. ICML'18], and Poincaré GloVe [Tifrea et al. ICLR'18].

Results

Model	Random Negatives			Hard Negatives		
	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

Strong performance but do not support inductive predictions

Not doing well on hard negatives but with high recall → good on predicting positives but not separating negatives

HiTs consistently perform better than all baselines

Support inductive predictions but limited by word-level vocabulary

Analysis

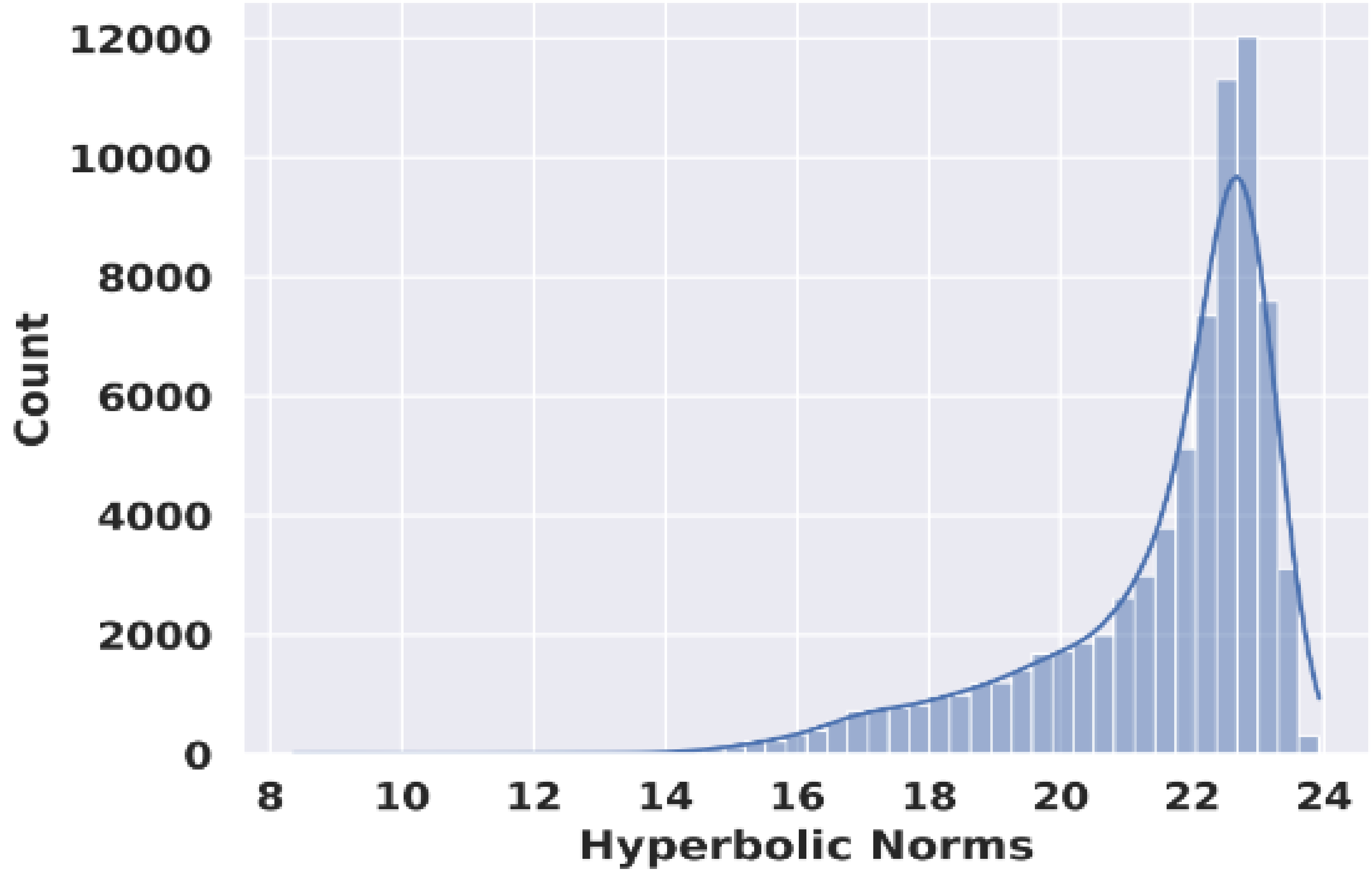


Fig. Distribution of entity hyperbolic norms.

HIT	PoincaréEmbed	HyperbolicCone
0.346	0.130	0.245

Table. Statistical correlations between entities' hyperbolic norms and depths in the WordNet hierarchy.

	computer	pc	fruit	berry
computer	0.0	5.9	22.5	24.9
pc	5.9	0.0	25.2	27.2
fruit	22.5	25.2	0.0	6.72
berry	24.9	27.2	6.72	0.0
h-norm	17.5	19.1	15.3	16.6
depth	9	11	9	10

Table. Case study of specific entity embeddings.

Future Work

- Encoding multiple hierarchical relationships within one model (**MHIT**?)
- Mitigate catastrophic forgetting resulted from hierarchy re-training
- Hierarchy-based semantic search that contrasts with traditional similarity-based one
- Pre-train, or further pre-train an LM on a large set of hierarchies

THANKS!

Contact

Dr. Yuan He	yuan.he@cs.ox.ac.uk
Zhangdie Yuan	zy317@cam.ac.uk
Dr. Jiaoyan Chen	jiaoyan.chen@manchester.ac.uk
Prof. Ian Horrocks	ian.horrocks@cs.ox.ac.uk

UNIVERSITY OF OXFORD

DEPARTMENT OF COMPUTER SCIENCE

NLP UNIVERSITY OF CAMBRIDGE

MANCHESTER 1824 The University of Manchester

NEURAL INFORMATION PROCESSING SYSTEMS