Exploring Supervised Learning of Hierarchical Event Embedding with Poincaré Embeddings

Pride Kavumba† Naoya Inoue†;‡ Kentaro Inui†;‡

†Tohoku University  ‡RIKEN Center for Advanced Intelligence Project (AIP)
kavumba.pride.q2@dc.tohoku.ac.jp
{inoue, inui}@ecei.tohoku.ac.jp

1 Introduction

Understanding events expressed in text is important in many natural language understanding tasks such as dialogue systems, question answering, discourse understanding, and information extraction. Natural language sentences and events exhibit hierarchical structure [8, 1]. This hierarchy can be defined in terms of specificity. General events can be considered as parent events to the more specific events. For instance, the event "person eats food" can be considered as the parent event to "John ate an apple" (see Figure 1 for more examples).

Capturing this kind of hierarchy is important in many applications such as causality recognition. For example, if we have the hierarchy for these events; "eat something", "eat food", and "eat apple". It is enough to only know that eat something causes someone to be full.

Previous work in event understanding [18, 23, 17, 9, 14, 5] use distributed representational of events in Euclidean space. In the recent work, it has been demonstrated that hyperbolic space is more suitable for learning representations for data which exhibit some kind of hierarchical structure such as nouns, social, semantic and complex networks [3, 15, 1, 12].

However, embedding events into non-Euclidean space has not yet been explored very well. Some previous work explored embedding words into other spaces to represent specificity of concepts [15]. Dhingra et al. [6] extend Nickel and Kiela’s 2017 [15] work to learn sentence encoder that can embed sentences into hyperbolic space by using an unsupervised Skip Thought-based objective [20]. However, the extrinsic evaluation, e.g. sentiment classification, results do not show significant improvement over Euclidean space-based encoder. In addition, they do not analyze the learned embeddings deeply. They try to learn hierarchical structure exhibited in the data implicitly. However, before going into fully unsupervised approaches, we believe that we should explore the properties of learned event embeddings with explicit supervision of hierarchical structure. Specifically we explore the following research questions:

1. Can hyperbolic embedding, proposed for non-structured entity (i.e. word), be adopted for structured entity (i.e. event)?
2. Does hyperbolic event embedding capture specificity of events?

We explore [15]-like approach, where the model is explicitly informed what concepts form a hierarchy. The contribution of our work can be summarized as follows:

• This is the first study to explore event embeddings learned with explicit supervision of event hierarchy.
• Our experiments demonstrate that hyperbolic event embeddings learned with explicit supervision capture the hierarchical nature of events.
• We show that, even without explicit supervision of word-based conceptual hierarchy, the learned embedding captures the hierarchy of words.
• We also show that learned hyperbolic event embeddings generalize well to unseen events.

2 Preliminaries

2.1 Word embeddings in hyperbolic space

Next, we describe word embeddings in hyperbolic space as presented by [15]. Hyperbolic geometry is the geometry you obtain by assuming all the postulates of Euclid, except the fifth one, which is replaced by its negation. That is, in hyperbolic geometry there exist a line $l$ and a point $P$ not on $l$ such that at least two distinct lines parallel to $l$ pass through $P$.

In a regular tree, the number of children at each node grows exponentially with the distance from the node.
hyperbolic space, the circumference and the area of the circle is proportional to the sinh of the radius and the cosh of the radius, respectively. This exponential relationship with the radius makes hyperbolic space more suitable for embedding trees. A regular tree can be easily embedded in 2-dimensional hyperbolic space. However, previous work [15, 21, 6], has shown that as the dimension of the hyperbolic embedding increases, the performance on downstream tasks improves.

Of the many hyperbolic space models, Poincaré ball model is the most suitable model for representational learning in neural networks as it is more suitable for gradient-based optimization. Poincaré ball model can be defined as unit ball, \(B^d = \{ x \in \mathbb{R}^d \| x \| < 1 \} \), were the distance between two points \( u \) and \( v \) is given by:

\[
d(u, v) = \arccosh \left( 1 + \frac{2}{1 - \| u \|^2} \frac{\| u - v \|^2}{1 - \| v \|^2} \right)
\]

To learn Poincaré embeddings, \( \Theta = \{ \theta_i \}_{i=1}^n \) for a set of symbols \( \mathcal{E} = \{ e_i \}_{i=1}^n \), we need to define an objective function, \( L(\Theta) \), that minimizes the hyperbolic distance (1) between embeddings of related events. Then, we need to optimize:

\[
\Theta' \leftarrow \arg \min_{\Theta} L(\Theta) \quad \text{s.t.} \quad \forall \theta_i \in \Theta : \| \theta_i \| < 1
\]

[15] optimizes this equation using Riemannian Stochastic Gradient Descent (RSGD) [2]. In this work, we use the re-parametrization technique proposed by [6], described in section 2.2. In RSGD-based optimization, used by [15], it is possible that some embeddings can lie outside the Poincaré ball. Therefore, it is necessary to project such embeddings back in the Poincaré ball during each update. However, with re-parameterization technique [6], the projection is not necessary as the resulting embedding vectors always lie in the Poincaré ball. As a result of this, we can make use of any available optimizer such as Adam [11]. In addition, it was shown that training using re-parameterization converges faster while offering comparable results, in a similar task setting, to the work by [15].

2.2 Parametric Poincaré Embedding

Given event \( e_i \) and its embedding \( e(s) \), we compute:

\[
\nabla = \phi_{\text{dir}}(e(s)), \quad \nu = \frac{\nabla}{\| \nabla \|},
\]

\[
\bar{p} = \phi_{\text{norm}}(e(s)), \quad p = \sigma(\bar{p}),
\]

where \( \phi_{\text{dir}} \) and \( \phi_{\text{norm}} \) are arbitrary parametric functions, whose parameters are learned during training. We then obtain hyperbolic embedding \( \theta = p \nu \).

3 Hyperbolic Event Embeddings

3.1 Model

We use the re-parametrization technique described in section 2.2. Our approach is event encoder-agnostic. For \( e(s) \), we can employ any kind of sentence encoder that outputs fixed-length vector. We use LSTM as an event encoder. For projections, we use the following parametric function: \( \phi_{\text{dir}}(x) = W_d^T x \), \( \phi_{\text{norm}}(x) = W_p^T x \) We expect that event embeddings will be organized in hierarchical manner such that more general events will appear closer to the origin and more specific events will appear towards the edge.

3.2 Training Objective

To learn representations \( \Theta = \{ \theta_i \}_{i=1}^n \) for a set of events \( \mathcal{E} = \{ e_i \}_{i=1}^n \), we define a loss function \( L(\Theta, d) \) that minimizes the hyperbolic distance (1) between embeddings of related events.

\[
L(\Theta, d) = \sum_{(u, v) \in \mathcal{D}} \log \frac{e^{-d(u, v)}}{\sum_{v' \in N(u)} e^{-d(u, v')}}
\]

where \( u \) and \( v \) are composition vectors from the sentence encoder. It is worth noting that because the hyperbolic distance is symmetrical, the loss function does not use any directions of edges between \( u \) and \( v \).

4 Experiments & Results

4.1 Training Data

In our experiment, we use part of the entailment datasets, (KS2016), introduced by Kartasakis and Sadrzadeh [10] which consist of 70 subject-verb-object (SVO) pairs, \((u, v)\) were \( v \) is a more general event of \( u \). From each event \( u \) in the dataset, we use WordNet hyponyms to get more specific events. In addition, we use WordNet hypernyms to get more general events.

Using the aforementioned method we get 12,803 positive SVO pairs, \((u, v)\), and a vocabulary of 6027 unique words. This dataset, \( \mathcal{D} \), is arranged in the form \( \mathcal{D} = \{(u, v) | v \text{ is the general event of } u \} \). For each positive SVO pair \((u, v)\), we generate negative example by pairing \( u \) with randomly sampled SVO triplet. We split the dataset into train/test in the ratio 4:1.

4.2 Model Settings

Our model consists of three layers. The first layer of our event encoder is an Embedding layer. This layer uses pretrained 100-dimensional glove vectors [16]. The second layer is an LSTM layer with 64 units, and the third and final layer is the Projection Layer. The projection layers projects to 128-dimensional Poincaré embedding space. We use the Adam optimizer [11] for optimization.
0.262 - 0.271 physical object accent abstract entity
0.271 - 0.279 organism see abstract entity, somebody show abstract entity, living thing determine abstract entity, abstract entity show abstract entity, abstract entity verbalise noesis, physical object transfer abstract entity
0.279 - 0.288 woman experience abstract entity, idea fit abstract entity, person accept abstract entity, group verbalize abstract entity
0.288 - 0.296 human action give abstract entity, event show abstract entity, physical object show cognition
0.296 - 0.305 physical object change abstract entity, artifact represent abstract entity, psychological feature cerebrate abstract entity, physical entity interact substance, psychological feature move abstract entity, animate thing obtain abstract entity

| Table 1: Events with the lowest norms |
|-----------------|-----------------|
| Norm            | Event            |
| 0.99621 - 0.99623 | police catch ripper, term tell law of reciprocal proportions, muta'azeen skirmish statutory offence |
| 0.99623 - 0.99626 | term tell first law of motion, police catch jailbird, police collect suffering, term tell third law of motion, acute glossitis indicate cause |
| 0.99626 - 0.99628 | flush indicate cause, police catch collaborator, dowager wholesale legal jointure, police torpedo crime |
| 0.99628 - 0.9963  | police lump evidence, fever indicate cause, police catch rustler |
| 0.9963 - 0.99632  | hypoglycemia indicate cause, term slip in concept, police catch stickup man |

4.3 Intrinsic Qualitative Evaluation

We compare the norms of events from the resulting embedding. More general events are supposed to have lower norms than more specific events because general events are embedded closer to the origin. The results of this experiment are show in table 1 and table 2. Additionally, figure 2 shows the visualization of the 2D Poincaré event Embedding. For visual clarity, we manually picked only a few events from the KS2016 dataset. The resulting embedding capture the hierarchical nature of events, it places more general events closer to the origin.

In addition, we also compare the word level norms. Words which express more general concepts are supposed to lie closer to the origin than words the express specific concepts. The results for this experiment are shown in table 3 and table 4. Figure 3, shows the visualization of the word embedding obtained from 2D Poincaré event Embedding. For visual clarity, we manually picked a few related words from the training set. Even without explicit supervision of word-based conceptual hierarchy, the learned embedding captures the hierarchy of words.

4.4 Intrinsic Quantitative Evaluation

In the dataset, $D$, for each positive pair, $P(u, v)$ and its negative counterpart $N(u, v')$, we calculate is-a score [15].

$$\text{score(is-a}(x, y)) = -(1 + \alpha(||y|| - ||x||))d(x, y) \quad (3)$$

| Table 2: Events with the largest norms |
|-----------------|-----------------|
| Norm            | Event            |
| 0.573 - 0.582   | fauna, vertebrate, abstract, art, entity, interact |
| 0.523 - 0.533   | cognition, department, europol, organism, host, aspect, landscape, intelligence, biological, nestle, reality, defense, complex, interior, index, germany |
| 0.553 - 0.573   | culture, raw, chemical, flora, abstraction, fleischer, agency, language, nature, food, speak, flavor, rubin, situation, affairs, kraft, agriculture |

5 Related Work

Research on event understanding ranges from inferring intent and emotional reaction [18], sentiment classification [7], and script knowledge [19] modeling [4, 9, 14, 5, 17, 23, 13]. Previous work in event understanding [18, 23, 17, 9, 14, 5] use distributed representational of events in Euclidean space. However, in recent work, it has been demonstrated that hyperbolic space is more suitable for learning representations for data which exhibit some kind of hierarchical structure.

A variety of approaches have been proposed to capture the hierarchical structure of datasets. Vilnis et al. [22] proposed Gaussian Embeddings to capture uncertainty and asymmetry. Nickel and Kiela [15] learned word embeddings on Poincaré’s ball, while our work focuses on event embeddings. Tay et al. [21] learned question and answer embeddings on the Poincaré’s ball for question-answer retrieval. Dhingra et al. [6] extended [15] work and [6] showed a method to embed words and sentences into hyperbolic space. However, the extrinsic evaluation results do not show significant improvement over Euclidean space-based encoder, and they do not analyze the learned embeddings deeply. They learned hierarchical structure exhibited in the data implicitly using an unsupervised Skip Thought-based objective [20]. However, before going into fully unsupervised approaches, we believe that we should explore the properties of learned event embeddings with explicit supervision of hierarchical structure.
6 Conclusion and future work

In this paper, we presented the first study that explores learning event embeddings with explicit supervision of event hierarchy. We demonstrated that hyperbolic event embeddings learned with explicit supervision capture the hierarchical nature of events. We also showed that, even without explicit supervision of word-based conceptual hierarchy, the learned embedding also captures the hierarchy of words. Finally, we showed that the learned hyperbolic event embeddings generalize well to unseen events.

In future we intend to perform extrinsic evaluation of the hyperbolic event embedding and to attempt script knowledge modeling. In addition, we intend to explore if conventional point-based embedding capture general-specific relations of events. We also intend to compare our method with Skip-Thought [6] objective. Finally, we intend to consider monotonicity for real-world situation like Multi-NLI.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 15H01702 and JST CREST Grant Number JPMJCR1513, including AIP challenge.

References