Hype-HAN: Hyperbolic Hierarchical Attention Network for Semantic Embedding

Chengkun Zhang , Junbin Gao

Discipline of Business Analytics, The University of Sydney Business School
The University of Sydney, Camperdown, NSW 2006, Australia
{chengkun.zhang, junbin.gao}@sydney.edu.au

Abstract

Hyperbolic space is a well-defined space with constant negative curvature. Recent research demonstrates its odds of capturing complex hierarchical structures with its exceptional high capacity and continuous tree-like properties. This paper bridges hyperbolic space's superiority to the power-law structure of documents by introducing a hyperbolic neural network architecture named *Hyperbolic Hierarchical Attention Network (Hype-HAN)*. Hype-HAN defines three levels of embeddings (word/sentence/document) and two layers of hyperbolic attention mechanism (word-to-sentence/sentence-to-document) on Riemannian geometries of the *Lorentz model*, *Klein model* and *Poincaré model*. Situated on the evolving embedding spaces, we utilize both conventional GRUs (Gated Recurrent Units) and hyperbolic GRUs with Möbius operations. Hype-HAN is applied to large scale datasets. The empirical experiments show the effectiveness of our method.

1 Introduction

Semantic embedding is one of the most fundamental tasks of natural language processing (NLP). It aims to represent textual elements with low-dimensional latent variables while preserving their discriminative features. We witness the explosive growth of the internet and the resulting enormous textual content at our disposal. It becomes an important task to develop efficient and effective methods to process such amount of textual information.

Automatic text classification is a ubiquitous application of semantic embedding. With the pre-assigned labels on the training documents, one can complete the task with numerous feature selection techniques and various classifying architecture in favour of different focuses. In terms of traditional approaches, people may encode raw documents as sparse lexical features, such as n-grams [Damashek, 1995], tf-idf [Ramos, 2003] and bag-of-words [Wallach, 2006]. These features can be sent into classical classifiers, such as linear model or kernel methods [Wang and Manning, 2012; Lodhi *et al.*, 2002]. Recently, as computational power develops, neural network based methods have obtained its momentum in the field.

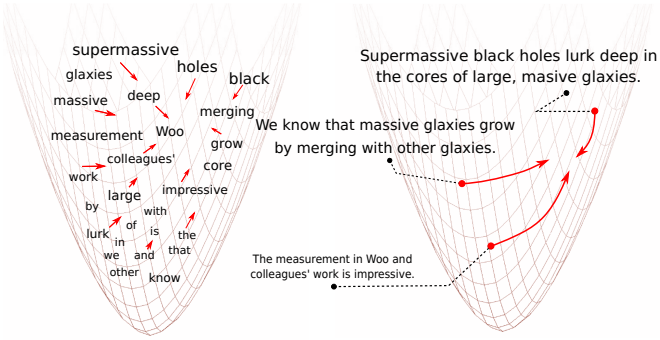


Figure 1: The hyperbolic embedding as a word-to-sentence and sentence-to-document knowledge accumulation process on a 2-dimensional Lorentz Model.

Typical practices include methods founded on convolutional neural network and recurrent neural network [Kim, 2014; Kim *et al.*, 2016; Yao *et al.*, 2019; Zhao *et al.*, 2019]. In particular, a number of researchers realize the natural hierarchical structure of documents and start designing specific neural network architectures acordingly [Yang *et al.*, 2016; Gao *et al.*, 2017; Ying *et al.*, 2018].

However, most of the main-stream methods are defined on Euclidean space, which is not particularly suitable or requiring tremendous embedding size for capturing complex networks. Though large embedding size may not be considered as a major drawback any longer thanks to the light-speed computational devices and the inundant memory, it is still a hard open question to interpret the hierarchical meaning of purely numerical latent variables. This paper tackles this gap by introducing a new network structure to embed the semantic relations on a self-informative manifold.

The hyperbolic manifold is a smooth manifold of constant negative curvature. It was studied in the ancient differential geometry [Cannon *et al.*, 1997] but just started gaining more attention in machine learning field lately [Nickel and Kiela, 2017; Nickel and Kiela, 2018; Ganea *et al.*, 2018; Gulcehre *et al.*, 2018]. With its constant negative curvature, the distance within this type of manifold is self-informative and the resulting space volume grows exponentially even in dimension of two. These characteristics make it particularly advisable for encoding complex networks.

Different from previous hyperbolic NLP practices with manually annotated pair-wise relation, such as [Nickel and Kiela, 2017; Nickel and Kiela, 2018], which consider the supervised word-to-word link hierarchy, or [Ganea *et al.*, 2018], which considers the sentence-to-sentence semantic relation, we regard word-to-sentence and sentence-to-document as two types of natural hierarchical labels. We build our work upon the well-known hierarchical attention architecture (HAN) [Yang *et al.*, 2016; Gao *et al.*, 2017; Cheng *et al.*, 2017; Xing *et al.*, 2018; Li *et al.*, 2018], which considers a knowledge accumulation process according to the structure of the target. We practice this idea and extend it in various ways:

- As complex networks require large embedding size and hierarchical relations are difficult to interpret in Euclidean space, we constrain the computational process in hyperbolic space to incorporate advantages of hyperbolic manifolds. Since hyperbolic coordinate is self-informative, this enables us to include not only the hierarchical structure of explicit semantic levels (e.g. between words and sentences), but also the implicit parent-child relation among the same level (e.g. among words).
- We encode semantic representations from sequences instead of pre-defined tokens through gated recurrent units (GRUs) [Cho *et al.*, 2014]. As the operational manifolds are evolving along the forward path, to remain the geometric features, we utilize Euclidean GRUs and Möbius GRUs accordingly. Readers can find a conceptual architecture in Figure 3.
- We believe it is ill-posed to allocate semantic elements in different spaces since sometimes a single word and the whole sentence can be equivalently meaningful. Instead, we design our architecture delicately to ensure representations are aggregated from their sub-trees via hyperbolic attention mechanism.
- We utilize isometric hyperbolic models considering their different geodesic features in favour of computational convenience[1]. Features of word sequences are firstly projected on the *Lorentz model*. Then, word representations are aggregated on the *Klein model* via *Einstein midpoint*. Similar procedure are conducted for sentence accumulation except that sentence encoder is defined on *Poincaré model* due to its convenient Möbius operations.

2 Hyperbolic Geometry Initiation

Hyperbolic space, specifically referred to manifolds with constant negative curvature in this paper, has been studied in differential geometry for long, considered under five isometric models [Cannon *et al.*, 1997]. Among them, *Poincaré ball model* , *Lorentz model* and *Klein model* have received increasing attention in machine learning community due to their attractive features for modeling complex network.

[1]Geometrical properties are fully kept when representations are transformed between isometric models. Readers could refer to [Ratcliffe *et al.*, 1994] for a more detailed explanation.

Poincaré ball model. The Poincaré ball model is defined as a Riemannian manifold $\mathcal{P}^n = (\mathcal{B}^n, g^{\mathcal{P}})$, where $\mathcal{B}^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| < 1\}$, and the metric tensor $g^{\mathcal{P}}(\mathbf{x}) = (\frac{2}{1-\|\mathbf{x}\|^2})^2 g^{\mathcal{E}}$, with the Euclidean metric tensor $g^{\mathcal{E}}$. The distance on this manifold is defined as:

$$d_p(\mathbf{x}, \mathbf{y}) = \text{arcCosh}\left(1 + 2\frac{\|\mathbf{x}-\mathbf{y}\|^2}{(1-\|\mathbf{x}\|^2)(1-\|\mathbf{y}\|^2)}\right). \quad (1)$$

Lorentz model. The Lorentz model is the only unbounded hyperbolic model and is defined as $\mathcal{L}^n = (\mathcal{H}^n, g^{\mathcal{L}})$ with points constrained by $\mathcal{H}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -1, x_0 > 0\}$, where $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}$ is the Lorentzian scalar product:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{n} x_n y_n,$$

and the metric tensor is: $g_{\mathcal{L}}(\mathbf{x}) = \text{diag}(-1, 1, ..., 1)$. The pairwise distance function is given as:

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \text{arcCosh}(-\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}).$$

Since all hyperbolic models are isometric, one can project data from Lorentzian coordinates to Poincaré ball by:

$$\mathcal{L}^n \to \mathcal{P}^n : \ \text{L2P}(x_0, \ldots, x_n) = \frac{(x_1, \ldots, x_n)}{x_0 + 1}. \quad (2)$$

Klein Model. Klein model is also defined on $\mathcal{K}^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| < 1\}$. The isomorphism between Klein model and Poincaré ball can be defined through a projection on or from the hemisphere model. People can get Klein points from Poincaré coordinates by:

$$\mathcal{P}^n \to \mathcal{K}^n : \ \text{P2K}(\mathbf{x}_{\mathcal{K}}) = \frac{2\mathbf{x}_{\mathcal{P}}}{1 + \|\mathbf{x}_{\mathcal{P}}\|^2}, \quad (3)$$

or from Lorentzian coordinates by:

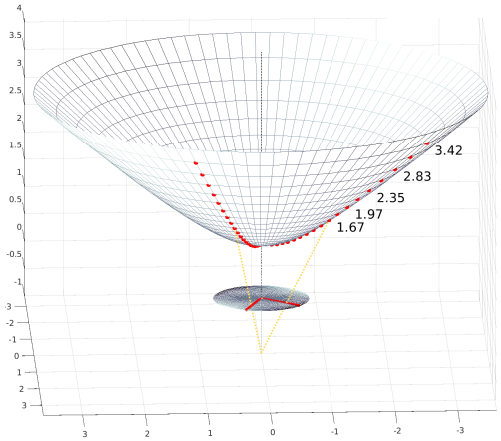
$$\mathcal{L}^n \to \mathcal{K}^n : \ \text{L2K}(x_0, \ldots, x_n) = \frac{(x_1, \ldots, x_n)}{x_0}. \quad (4)$$

3 Hyperbolic Hierarchical Attention Network

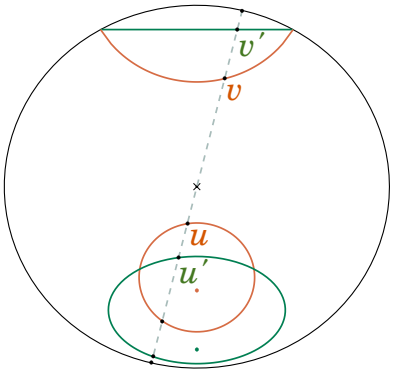
In the following, we construct a knowledge accumulation process based on hyperbolic geometry and utilize the attention mechanism to capture the semantic subordination. We treat word-to-sentence and sentence-to-document as two natural hierarchical labels. We enforce latent variables of these semantic elements lie on a hyperbolic space and aggregate them according to a child-parent structure. An abstract architecture of Hype-HAN is demonstrated in Figure 3.

3.1 Hyperbolic Activation

Though conventional RNN-based method has proven its success in modeling sequence data, it suffers from a crucial limitation that the ability to model complicated hierarchies is tightly bounded by the dimensions of latent variables. Meanwhile, the hierarchical relation between components on the same semantic level (word-to-word or sentence-to-sentence) is not delicately taken care of since Euclidean space is not suitable to capture the existence of such power-law distribution. Thus, to utilize the elegant feature of hyperbolic models, we consider constraining latent variables on a Lorentz model.



(a) Projection between Lorentz model and Poincaré disk



(b) Projection between Poincaré model (red) and Klein model (green)

Figure 2: (a) Points on a unit Poincaré disk with in-between distances: $d_{\mathcal{P}} = 0.1$ are projected on the Lorentz model. The corresponding Lorentz factor $\gamma(\mathbf{x}_{\mathcal{K}}) = \frac{1}{\sqrt{1-\|\mathbf{x}_{\mathcal{K}}\|^2}}$ grows when points are closer to the boundary. (b) (i) The geodesics in a Poincaré disk are circular arcs perpendicular to the boundary (red) while the same line in a Klein model is straight (green) . (ii) Hyperbolic circles have their own hyperbolic radius. A circle in a Poincaré disk (red) is equivalent to an ellipse (green) in Klein model. But both centers are not at the Euclidean centers of the circle or the ellipse.

Denote the state output of word encoder as $\mathbf{h}_{it}$, we first transform this vector into its polarity form as $(\mathbf{d}, r) \in \mathbb{R}^{n+1}$, where $r = \|\mathbf{h}_{it}\|$ and $\mathbf{d} = \frac{\mathbf{h}_{it}}{r}$, i.e. $\|\mathbf{d}\| = 1$. Then, the following activation function is conducted to constrain the representation into a valid hyperbolic form:

$$\mathcal{E} \longrightarrow \mathcal{L} : \text{act}((\mathbf{d}, r)) = (\sinh(r)\mathbf{d}, \cosh(r)). \quad (5)$$

We can verify the validity of $\mathbf{h}_{t\mathcal{L}} = (\sinh(r)\mathbf{d}, \cosh(r))$ by checking whether it possesses Lorentzian characteristics from two aspects:

- Each point must have its squared Lorentz norm equal to -1. In our case, it is easy to show $\langle \mathbf{h}_{t\mathcal{L}}, \mathbf{h}_{t\mathcal{L}} \rangle_{\mathcal{L}} = -1$.
- The volume as embedding space should grow exponentially in terms of a linear increase in radius. In our case,

$$d_{\mathcal{L}}(\mathbf{0}, (\mathbf{d}, r)) = \text{arcCosh}(-\sinh(0)\sinh(r)\langle \mathbf{0}, \mathbf{d}_j \rangle + \cosh(0)\cosh(r)) = r,$$

  which also satisfies the condition. This en-powers the network with much larger capacity compared to Euclidean space for modeling sophisticated relations.

By using such activation technique, we have constrained the output of GRUs as hyperbolic points before sending them to the next layer. Since hyperbolic manifold has a continuous tree structure, it enables the components on the same layer of neural network to also share a hierarchical structure (among words or among sentences). Thus, the simple discrete double-level hierarchies (word-to-sentence and sentence-to-document) are en-powered as its continuous analogue.

3.2 Semantic Encoders

Figure 3 shows both word and sentence encoders based on GRUs to capture the semantic representation of sequences. The intuition of such design is considering that not only different words/sentences have different meaning, same words may also be differentially informative in different context, i.e. the sequence of words plays a significant role in semantic analysis.

Consider the embedding of word-t appeared in sentence-i as $\mathbf{x}_{it}$, which can be initialized from a linear layer or pre-trained vectors, the hidden state of $\mathbf{x}_{it}$ within the sentence can be constructed with a GRUs layer as:

$$\mathbf{h}_{it} = [\overrightarrow{\text{GRU}}(\mathbf{x}_{it}), \overleftarrow{\text{GRU}}(\mathbf{x}_{it})],$$

where $\overrightarrow{\text{GRU}}(\mathbf{x}_{it})$ learns the forward hidden state while $\overleftarrow{\text{GRU}}(\mathbf{x}_{it})$ learns the backward hidden state around the queried word.

Given that latent variables may lie on different manifolds (Euclidean or hyperbolic), GRU architectures used in this work are divided into two streams:

Euclidean GRU Architecture. With slight abuse of notation in this and the following paragraph, we denote input vector at time-t as $\mathbf{x}_t$, one can adapt the Euclidean GRU architecture as:

$$\mathbf{z}_t = \sigma(\mathbf{U}_z\mathbf{x}_t + \mathbf{W}_z\mathbf{h}_{t-1} + \mathbf{b}_z),$$
$$\mathbf{r}_t = \sigma(\mathbf{U}_r\mathbf{x}_t + \mathbf{W}_r\mathbf{h}_{t-1} + \mathbf{b}_r),$$
$$\tilde{\mathbf{h}}_t = \varphi(\mathbf{U}_h\mathbf{x}_t + \mathbf{W}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h),$$
$$\mathbf{h}_t = \mathbf{z}_t \odot \tilde{\mathbf{h}}_t + (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1},$$

where σ and φ are point-wise non-linearity, $\odot$ is point-wise product, $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{b}$ are parameters, $\tilde{\mathbf{h}}_t$ is the new state computed from the new sequence, $\mathbf{r}_t$ decides how much information should be preserved to construct $\tilde{\mathbf{h}}_t$, while $\mathbf{z}_t$ decides the contribution proportion of the past and new information to construct $\mathbf{h}_t$.

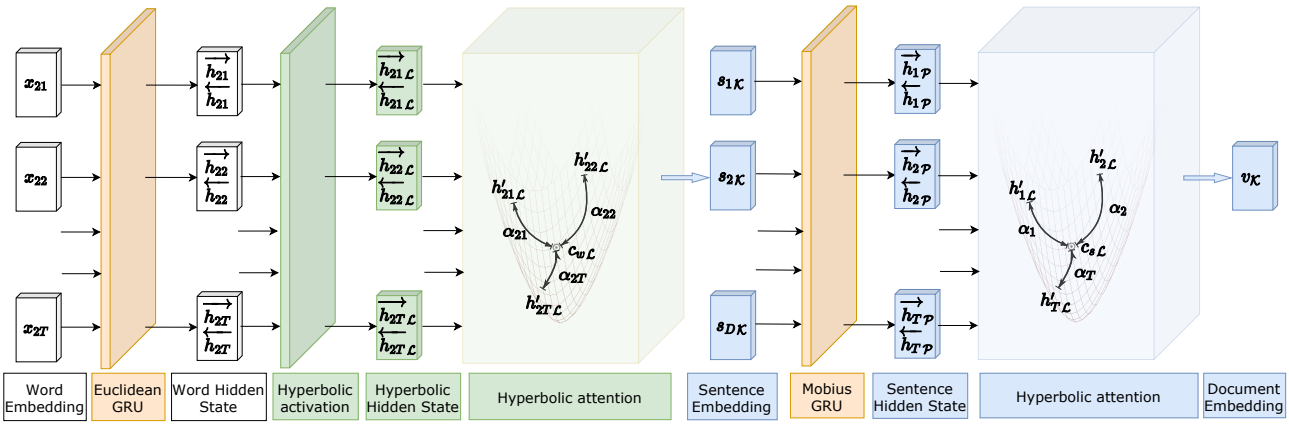


Figure 3: An Abstract Architecture of Hyperbolic Hierarchical Attention Network

However, as can be noticed, Euclidean GRU is based on several linear layers with non-linearity manually added. This is not valid on manifold spaces, thus we employ hyperbolic GRUs after latent variables are projected on hyperbolic space.

Hyperbolic GRU Architecture. Ganea [2018] introduced Möbius operations on Poincaré model, including Möbius addition, multiplication and bias, which are the key elements a GRU needs. Thus, Möbius GRU on the Poincaré model is defined as:

$$\mathbf{z}_t = \sigma \log_0(((\mathbf{W}_z \otimes \mathbf{h}_{t-1}) \oplus (\mathbf{U}_z \otimes \mathbf{x}_t)) \oplus \mathbf{b}_z),$$
$$\mathbf{r}_t = \sigma \log_0(((\mathbf{W}_r \otimes \mathbf{h}_{t-1}) \oplus (\mathbf{U}_r \otimes \mathbf{x}_t)) \oplus \mathbf{b}_r,$$
$$\tilde{\mathbf{h}}_t = \varphi^{\otimes}((([\mathbf{W}_h \mathrm{diag}(\mathbf{r}_t)] \otimes \mathbf{h}_{t-1}) \oplus (\mathbf{U}_h \otimes \mathbf{x}_t)) \oplus \mathbf{b}_h),$$
$$\mathbf{h}_t = \mathbf{h}_{t-1} \oplus (\mathrm{diag}(\mathbf{z}_t) \otimes ((-\mathbf{h}_{t-1}) \oplus \tilde{\mathbf{h}_t})),$$

where $\log_0$ is the logarithmic map, $\varphi^{\otimes}$ is the Möbius non-linearity and $\mathrm{diag}(u)$ is the square diagonal matrix.

The rationale of using different types of GRU architecture is to utmostly remain the geometric features of semantic embeddings. With such design, different semantic elements at different level can all utilize the geodesic features of hyperbolic manifolds.

3.3 Hyperbolic Semantic Aggregation

We now discuss how to conduct the ubiquitous attention mechanism to aggregate hyperbolic individual components.

Word-to-Sentence Aggregation

As discussed in the previous, we activate hidden state of words onto hyperbolic space. It is obvious that words would not contribute equally to the meaning of one sentence. Thus, we need to construct a meaningful aggregation strategy.

Different from Euclidean space with constant zero curvature, hyperbolic attentions are not very obvious to design. Inspired by [Ganea *et al.*, 2018], we exercise the idea of using Einstein midpoint to compute the aggregation weights of attention and demonstrate new interpretation in the context of textual analysis.

Consider the hyperbolic hidden state $\mathbf{h}_{it\mathcal{L}}$, in order to utilize the Einstein midpoint defined on Klein model, we first take the transformation (4): $\mathbf{h}_{it\mathcal{L}} \rightarrow \mathbf{h}_{it\mathcal{K}}$. Then, to construct the attention weights for aggregation, we utilize the hyperbolic power-law characteristics to measure the importance of individual components. We aim to jointly learn a hyperbolic word centroid $\mathbf{c}_{w\mathcal{L}}$ from all the training documents. $\mathbf{c}_{w\mathcal{L}}$ can be considered as a baseline for measuring the importance of hyperbolic words based on their mutual distance. The underlying reasoning of such design generally considers two aspects: First, though training documents have been labeled independently, certain hidden states should play a dominant role. Second, traditional documentation methods consider document labels as separate entities, but there may also exist principal and subordinate structure among document labels. For instance, in terms of research paper classification, documents may be labeled as 'Topology', 'Mathematics', 'Social Science' and 'Psychology'. Obviously, 'Mathematics' and 'Topology' may abide by certain hierarchical relation while the same for 'Social Science' and 'Psychology'.

To learn $\mathbf{c}_{w\mathcal{L}}$, we consider another layer upon hidden state $\mathbf{h}_{it}$ as:

$$\mathbf{h}_{it}{}' = \tanh(\mathbf{W}_w \mathbf{h}_{it} + \mathbf{b}_w). \tag{6}$$

Similar to projecting $\mathbf{h}_{it}$ onto hyperbolic space, we use (5) to activate $\mathbf{h}_{it}{}'$ as $\mathbf{h}_{it\mathcal{L}}'$. Then the attention weights is computed as α_{it} by:

$$\alpha_{it} = \exp(-\beta_w d_{\mathcal{L}}(\mathbf{c}_{w\mathcal{L}}, \mathbf{h}_t{}') - c_w). \tag{7}$$

After capturing the hyperbolic attention weights, semantic meaning of words appearing in the same sentences has been aggregated via Einstein midpoint:

$$\mathbf{s}_i = \sum_t \left[\frac{\alpha_{it}\gamma(\mathbf{h}_{it\mathcal{K}})}{\sum_l \alpha_{il}\gamma(\mathbf{h}_{it\mathcal{K}})} \right] \mathbf{h}_{it\mathcal{K}}, \tag{8}$$

where $\gamma(\mathbf{h}_{it\mathcal{K}})$ is the so-called Lorentz factors, which can be expressed as:

$$\gamma(\mathbf{h}_{it\mathcal{K}}) = \frac{1}{\sqrt{1 - \|\mathbf{h}_{it\mathcal{K}}\|^2}} = \frac{1}{\sqrt{1 - \frac{\sinh^2(r_{it})}{\cosh^2(r_{it})}}}. \tag{9}$$

Since Klein and Poincaré ball model reflect the power-law distribution with its radius, it is not difficult to interpret the above function as a relativistic importance factor.

Sentence-to-Document Aggregation

Similar to the word-to-sentence attention, sentences with more relevant information of the topic can be the clues to classify the documents, thus we use the similar attention mechanism to aggregate sentences as document representation.

However, as we mentioned in the previous that the operational space evolved from Euclidean to hyperbolic manifold, the linear layer (6) will no longer hold on a curved manifold. Thus, we consider a 'linear' layer within Poincaré space:

$$\mathbf{h}_i{}'_{\mathcal{P}} = \mathbf{W}_{s\mathcal{P}} \otimes \mathbf{h}_{i\mathcal{P}} \oplus \mathbf{b}_{s\mathcal{P}}. \tag{10}$$

No $\tanh()$ is applied because this layer has already preserved non-linearity.

We first take: $\mathbf{h}_{i\mathcal{P}} \rightarrow \mathbf{h}_{i\mathcal{K}}$ and $\mathbf{h}'_{i\mathcal{P}} \rightarrow \mathbf{h}'_{i\mathcal{L}}$ via (3) and the reversed (2), then construct the document representation via Einstein midpoint:

$$\alpha_i = \exp(-\beta_s d_{\mathcal{L}}(\mathbf{c}_{s\mathcal{L}}, \mathbf{h}'_{i\mathcal{L}}) - c_s), \tag{11}$$

$$\mathbf{v}_{\mathcal{K}} = \sum_t \left[\frac{\alpha_i \gamma(\mathbf{h}_{i\mathcal{K}})}{\sum_l \alpha_l \gamma(\mathbf{h}_{i\mathcal{K}})} \right] \mathbf{h}_{i\mathcal{K}}, \tag{12}$$

where $\mathbf{c}_{s\mathcal{L}}$ is the sentence centroid, which can be randomly initialized and jointly learned with the other network parameters, and $\mathbf{v}_{\mathcal{K}}$ is the semantic meaning of the whole document summarized from all sentences.

Since $\mathbf{v}_{\mathcal{K}}$ is a high level representation of the whole document, similar to any other classification task, it can be sent to a hyperbolic 'linear layer' with softmax activation for document classification.

4 Evaluation

With the help of the self-informative and high-capacity hyperbolic manifolds, the proposed method brings forward its main advantages on explainability compared to the other neural network-based practices. We validate Hype-HAN from two aspects: (i) a conventional document classifier, to classify any large-scale datasets with performance on par with the state-of-the-art; and (ii) the core of a hierarchical interpreter or visualization toolbox, to demonstrate the hierarchical structure of the dataset on a tiny dimension through its hyperbolic coordinates.

4.1 Benchmark Comparison

In order to validate our work in the real-world scenario, we test Hype-HAN on some publicly available large-scale benchmark datasets with the same protocols from [Zhang *et al.*, 2015; Yogatama *et al.*, 2017]. The datasets include news classification (AGnews), question/answer categorization (Yahoo Answers), sentiment analysis (Yelp and Amazon) and Wikipedia article classification (DBpedia). The descriptive statistics of the dataset has been shown in table 1.

	Train	Test	Classes	Sent_Len	Doc_Len	Type
AG.	120,000	7,600	4	58	3	news
DBpedia	560,000	70,000	14	40	4	ontology
Yelp	600,000	7,600	2	33	15	reviews
Yahoo Answers	1,400,000	60,000	10	33	10	q&a
Amazon	3,600,000	400,000	2	33	9	reviews

Table 1: Descriptive statistics of datasets.

	AG.	DB.	Ye.	Yh.	Az.
N. Bayes	90.0	96.0	86.0	68.7	-
D-LSTM	92.1	**98.7**	92.6	**73.7**	-
G-LSTM	90.6	95.4	88.2	69.3	-
BoW	88.8	96.6	92.2	68.9	90.4
ngrams	92.0	98.6	**95.6**	68.5	92.0
n-TFIDF	**92.4**	**98.7**	95.4	68.5	91.5
ch-CNN	87.2	98.3	94.7	71.2	**94.5**
ch-CRNN	91.4	98.6	94.5	71.7	94.1
VDCNN	91.3	**98.7**	**95.7**	**73.4**	**95.7**
fastText	91.5	98.1	93.8	72.0	91.2
fastText(b)	**92.5**	98.6	**95.7**	72.3	**94.6**
HAN	91.5	98.1	93.3	72.3	93.2
Hype-HAN	**92.2**	**98.7**	94.5	**72.6**	94.1

Table 2: Test accuracy [%] on classification datasets. We run Hype-HAN with the same hyperparameters for all datasets. For all baselines, we show the best-reported numbers without augmentation.

We consider the best reported benchmarks as baselines, including: N.Bayes and LSTM from [Yogatama *et al.*, 2017], Bow, n-grams, n-TFIDF and ch-CNN from [Zhang *et al.*, 2015], ch-CRNN from [Xiao and Cho, 2016], VDCNN from [Conneau *et al.*, 2017], fastText from [Joulin *et al.*, 2017] and HAN from [Yang *et al.*, 2016].

To assess the generalization ability, we evaluate Hype-HAN with the same setting on different datasets. We initialize the word embedding via 'glove-50' [Pennington *et al.*, 2014], and we set the word/sentence hidden state as 50-dimension and train the models with manifold-aware Riemannian ADAM [Bécigneul and Ganea, 2018] with learning rate 0.001. We record the prediction accuracy on the test set around epoch 10-20 on smaller datasets (AGnews, DBpedia, Yelp), and around epoch 5-10 on the large-scale datasets (Yahoo and Amazon). The results are reported in Table 2.

It can be indicated that Hype-HAN outperforms its Euclidean ancestor on all datasets and achieves similar results with the state-of-the-art, especially on datasets related to topic classification, including AGnews, DBpedia, Yahoo answers. However, the performance on sentiment datasets (Yelp and Amazon reviews) is weaker compared to CNN-based methods and fastText(b) with bigram input. We suggest the phenomenon can be explained from several aspects:

- First, sentiment elements may share fewer hierarchical structure compared to ontology and taxonomy, which have straightforward power-law distributions. We conclude this as the major reason for the performance downturn in terms of predicting review ratings.
- Second, if the hierarchical information does not play a dominant role in the dataset, the depth of the network will account for more credits, which explains the advantages of CNN-based methods, such as the Very Deep Convolutional Networks [Conneau *et al.*, 2017].
- Third, as can be noticed, fastText(b) with bigram information is very robust while the version without utilizing bigram information has a large drop on the performance, especially on the sentiment tasks. As discussed by Joulin [2017], adding the bigram information on their architecture has increased the performance by 1-4%. We sug-

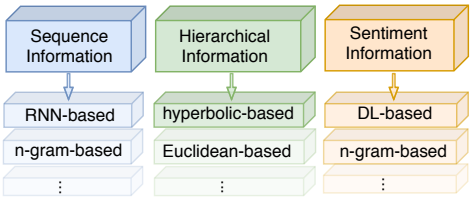


Figure 4: Priority of methods with different focuses summarized from the classification result.

gest that using bigram(n-grams) could also capture the sentimental information similar to CNN-based methods. Thus, for future work, one could investigate whether Hype-HAN could benefit from such extra information since it has been reported that RNN-based LSTM cell with dropout could be the best model for encoding the n-gram state [Chelba *et al.*, 2017].

4.2 Hyperbolic Interpretation

More importantly, compared to the other neural network-based classifier, Hype-HAN brings forward its main advantages on explainability, thus the usage of the proposed method is not limited for traditional classification tasks. Benefit from the self-informative hyperbolic coordinates, the learned semantic representations can be utilized directly for generating insights.

We scrape 1000 scientific abstracts from arXiv.org with keywords: 'manifold' and 'supermassive black hole' respectively. We split 2000 samples into 1600/400 training-test set and use only three-dimensional embedding space for both Euclidean and Möbius GRU to visualize the knowledge accumulation process. This experiment is conducted without a GPU and the embedding converges in 11 epochs with 99.7% prediction accuracy (default hyperparameter setting with the benchmark experiment).

We record the Poincaré hidden states of a random mini-batch at epoch 10 and visualize the points at the word, sentence and document level. As shown in Figure 5, the embedding of different semantic elements can be well-represented in the same hyperbolic space, which validates our hypothesis. Meanwhile, one could notice that the attention mechanism, based on Einstein midpoint (8) and (12), plays a prominent role in the aggregation step. Since the attention is computed based on Klein coordinates (with straight geodesics), the direction of the vector would dominate the process and motivate embeddings to evolve along separate directions.

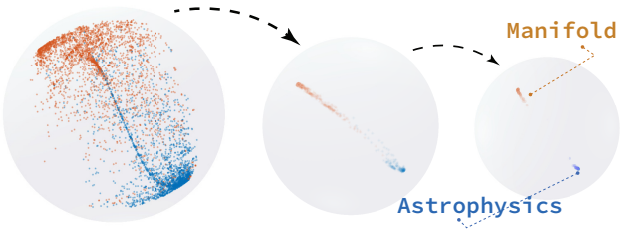
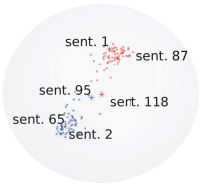


Figure 5: From the left to the right: the knowledge accumulation process at the word, sentence and document level.



Sent. 87: we prove conforml positive mass theorems for asymptotically flat manifolds with charge.
Sent. 65: pbhs much larger than this might have important consequences even if they provide only a small faction of the dark matter.
Sent. 118: We study the interaction between the geodesic orbit (g.o.~ in short) property and certain flag curvature condition.
Sent. 95: outflow size-luminosity relationenergetic gas outflows driven by active galactic nuclei (agns) are considered as one of the mechanisms.

Figure 6: Sentence embeddings from one batch during the training process. It could be noticed that sentences with keywords: 'manifolds' and 'dark matter' have a larger Lorentz factor.

To demonstrate the hierarchical structure of the semantic embeddings, we randomly extract the sentence representations from a mini-batch and visualize it in a Poincaré ball. Different from Euclidean variables, one could investigate the importance of semantic elements based on Lorentz factor (or radius). As shown in Figure 6, sentences with keywords: 'manifolds' and 'dark matter', are closer to the boundary of the defined space. Thus, apart from the traditional classifier usage, we suggest that one could utilize hyperbolic embeddings for more insights generation tasks, such as (i) low-dimensional representation for hierarchical visualization, and (ii) statistical methods to summarize the power-law distributions.

5 Related Work

Hierarchical structures are important in various of representation and logical reasoning tasks. Researchers reveal that perceptual representation in human brain may have a hierarchical relation [Palmer, 1977] and there is also physiological evidence supporting the parts-based representation theory in object recognition [Ullman, 1996]. People are inspired from such intuition and build practical algorithms. For instance, Lee and Seung [1999] use non-negative matrix factorization to represent images based on their partial information. Hyperbolic geometry, receives growing attention in hierarchical network embedding and machine learning field recently. For instance, Krioukov *et al.* [2010] demonstrates that heterogeneous degree distributions and strong clustering of complex network can be modeled with hyperbolic geometry. Nickel and Kiela [2017; 2018] discuss the supervised word-to-word link prediction tasks, and Ganea *et al.* [2018] considers the sentence-to-sentence textual entailment in hyperbolic space. We are inspired from these works and reformulate the problem as a manifold-based hierarchical knowledge accumulation process.

6 Conclusion

In this work, we practice a hierarchical embedding method: Hype-HAN based on three types of hyperbolic manifolds. The proposed method is evaluated with large-scale datasets for text classification tasks. Hype-HAN can be used as (i) the core of a hierarchical interpreter or visualizer without explicitly storing the discrete tree structure, or (2) a conventional document classifier with its performance on par with the state-of-the-art.

References

[Bécigneul and Ganea, 2018] G. Bécigneul and O.-E. Ganea. Riemannian adaptive optimization methods. *ICLR*, 2018.

[Cannon *et al.*, 1997] J. W. Cannon, W. J. Floyd, R. Kenyon, and W. R. Parry. Hyperbolic geometry. *Flavors of Geometry*, 1997.

[Chelba *et al.*, 2017] C. Chelba, M. Norouzi, and S. Bengio. N-gram language modeling using recurrent neural network estimation. *arXiv:1703.10724*, 2017.

[Cheng *et al.*, 2017] J. Cheng, S. Zhao, J. Zhang, I. King, X. Zhang, and H. Wang. Aspect-level sentiment classification with heat (hierarchical attention) network. In *CIKM*, pages 97–106, 2017.

[Cho *et al.*, 2014] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Conneau *et al.*, 2017] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *EACL*, pages 1107–1116, 2017.

[Damashek, 1995] M. Damashek. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843–848, 1995.

[Ganea *et al.*, 2018] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. In *NIPS*, 2018.

[Gao *et al.*, 2017] S. Gao, M. Young, J. Qiu, H. Yoon, J. Christian, P. Fearn, G. Tourassi, and A. Ramanthan. Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Inform Assoc.*, 25(3):321–330, 2017.

[Gulcehre *et al.*, 2018] C. Gulcehre, M. Denil, M. Malinowski, A. Razavi, R. Pascanu, K. Mo. Hermann, P. Battaglia, V. Bapst, D. Raposo, A. Santoro, and N. de Freitas. Hyperbolic attention networks. *ICLR*, 2018.

[Joulin *et al.*, 2017] A. Joulin, É. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *EACL*, pages 427–431, 2017.

[Kim *et al.*, 2016] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. In *AAAI*, pages 2741–2749, 2016.

[Kim, 2014] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751, 2014.

[Krioukov *et al.*, 2010] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

[Lee and Seung, 1999] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Li *et al.*, 2018] Z. Li, Y. Wei, Y. Zhang, and Q. Yang. Hierarchical attention transfer network for cross-domain sentiment classification. In *AAAI*, 2018.

[Lodhi *et al.*, 2002] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *JMLR*, 2(Feb):419–444, 2002.

[Nickel and Kiela, 2017] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *NIPS*, pages 6338–6347, 2017.

[Nickel and Kiela, 2018] M. Nickel and D. Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, pages 3779–3788, 2018.

[Palmer, 1977] S. E. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9(4):441–474, 1977.

[Pennington *et al.*, 2014] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Ramos, 2003] J. Ramos. Using tf-idf to determine word relevance in document queries. In *ICML*, 2003.

[Ratcliffe *et al.*, 1994] J. G. Ratcliffe, S. Axler, and K. Ribet. *Foundations of Hyperbolic Manifolds*. Springer, 1994.

[Ullman, 1996] S. Ullman. *High-level vision: Object Recognition and Visual Cognition*, volume 2. MIT Press Cambridge, MA, 1996.

[Wallach, 2006] H. M. Wallach. Topic modeling: beyond bag-of-words. In *ICML*, pages 977–984, 2006.

[Wang and Manning, 2012] S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL*, pages 90–94, 2012.

[Xiao and Cho, 2016] Y. Xiao and K. Cho. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv:1602.00367*, 2016.

[Xing *et al.*, 2018] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou. Hierarchical recurrent attention network for response generation. In *AAAI*, 2018.

[Yang *et al.*, 2016] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *AACL*, 2016.

[Yao *et al.*, 2019] L. Yao, C. Mao, and Y. Luo. Graph convolutional networks for text classification. In *AAAI*, volume 33, pages 7370–7377, 2019.

[Ying *et al.*, 2018] H. Ying, F. Zhuang, F. Zhang, Y. Liu, G. Xu, X. Xie, H. Xiong, and J. Wu. Sequential recommender system based on hierarchical attention networks. In *IJCAI*, pages 3926–3932, 2018.

[Yogatama *et al.*, 2017] D. Yogatama, C. Dyer, W. Ling, and P. Blunsom. Generative and discriminative text classification with recurrent neural networks. In *ICML*, 2017.

[Zhang *et al.*, 2015] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *NIPS*, pages 649–657, 2015.

[Zhao *et al.*, 2019] Y. Zhao, Y. Shen, and J. Yao. Recurrent neural network for text classification with hierarchical multiscale dense connections. In *IJCAI*, pages 5450–5456, 2019.