Joint Learning of Hyperbolic Label Embeddings for Hierarchical Multi-label Classification - HiddeN

Soumya Chatterjee[1] Ayush Maheshwari[1] Ganesh Ramakrishnan[1] Saketha Nath Jagarlapudi[2]

[1]{soumya, ayusham, ganesh}@cse.iitb.ac.in, [2]saketha@iith.ac.in

[1] Indian Institute of Technology Bombay [2] Indian Institute of Technology Hyderabad

Problem Statement

Given a set of documents and labels, classify the documents into multiple labels respecting the hierarchy. For eg., *Voice Recognition Is Improving, but Don't Stop the Elocution Lessons* - Labels are *Top/News/Technology*.

Assumption : Label hierarchy is not available.

Key Contributions

1. Our approach, HiddeN learns label embeddings using the joint optimisation approach
2. HiddeN sometimes generalizes even better than state-of-the-art hierarchical multi-label classifiers that have complete access to the true label hierarchy
3. We show significant improvement over classical multi-label classification methods as well as baselines that employ hyperbolic label embeddings.

Background: Poincaré Embeddings

- Let $\mathcal{B}^n = \{x \in \mathbb{R}^n | \; \|x\| < 1\}$ be the *open* n-dimensional unit ball, where $\|.\|$ is the Euclidean 2 norm.
- The Poincaré ball model is a Riemannian Manifold $(\mathcal{B}^n, g_x)$, the open unit ball equipped with the Riemannian metric tensor $g_x = \left(\frac{2}{1-\|x\|^2}\right)^2 g^E$, where $x \in \mathcal{B}^d$ and g^E is the Euclidean metric tensor.
- The geodesic distance between two points $u, v \in \mathcal{B}^d$ is given as

$$d(u,v) = arcosh\left(1+2\frac{\|u-v\|^2}{(1-\|u\|^2)(1-\|v\|^2)}\right)$$

Source Code

```
https://github.
com/soumyac1999/
hyperbolic-label-emb-for-hmc/
```

Our Model : HiddeN

- L are nodes of a fixed hierarchy but hierarchy is unknown to our model.
- **Document Model** - $\mathcal{F}_w(D) \in \mathbb{R}^n$
- **Label Embedding Model** - $\mathcal{G}_\Theta(l) \equiv \Theta * y^l = \Theta_l$, , where $\Theta \in \mathbb{R}^{n \times L}$
- Projection of $\Theta(l)$ into Poincaré manifold to get $\Pi(\Theta_l)$
  $\Pi(x) = \frac{x}{1+\sqrt{1+\|x\|_2^2}}$
- **Alignment Model:** $\hat{y}_D^l(w, \Theta) \equiv \sigma\left(\mathcal{F}_w(D)^\top \Theta_l\right)$

Joint Learning

- **First Term (Cross Entropy Loss for Classification)** -

$$\mathcal{L}_1(w,\Theta) = \sum_{i=1}^{m}\sum_{l=1}^{L}\left[y_i^l \log\left(\hat{y}_i^l(w,\Theta)\right) + (1-y_i^l)\log\left(1-\hat{y}_i^l(w,\Theta)\right)\right]$$

- **Second Term (Geodesic Distance Loss for Label Embeddings)** -

$$\mathcal{L}_2(\Theta) = \sum_{\substack{l,l' \in L,\\ l' \neq l}} \log\left(\frac{e^{-d(\Pi(\Theta_l),\Pi(\Theta_{l'}))}}{\sum_{z\in(l,l')} e^{-d(\Pi(\Theta_l),\Pi(\Theta_{l'}))}}\right)$$

- **Overall objective function**

$$\mathcal{L}(w,\Theta) = \mathcal{L}_1(w,\Theta) + \lambda\mathcal{L}_2(\Theta) \tag{1}$$

- **Inference:** Labels with $\hat{y}_D^l(\hat{w}, \hat{\Theta}) > 0.5$

Variants of HiddeN

1. $\text{HiddeN}_{\text{jnt}}$ - $(w_{\text{jnt}}, \Theta_{\text{jnt}}) \in \arg\min_{w,\Theta} \mathcal{L}(w,\Theta)$
2. $\text{HiddeN}_{\text{cas}}$
   - $\mathcal{L}_2$ is minimized to obtain label embeddings $\hat{\Theta}_{\text{cas}} \in \arg\min_\Theta \mathcal{L}_2(\Theta)$.
   - These are then used in $\mathcal{L}_1$ to obtain document parameters: $\hat{w}_{\text{cas}} \in \arg\min_w \mathcal{L}_1(w, \hat{\Theta}_{\text{cas}})$.
3. $\text{HiddeN}_{\text{flt}}$ - Θ_{flat} is fixed to the identity matrix
4. $\text{HiddeN}_{\text{euc}}$ $\mathcal{L}_{2\text{Euc}}(\Theta) = \sum_{\substack{l,l' \in L,\\ l' \neq l}} \log\left(\frac{e^{-\|\Theta_l-\Theta_{l'}\|_2}}{\sum_{z\in(l,l')} e^{-\|\Theta_l-\Theta_{l'}\|_2}}\right)$

Synthetic Experiments

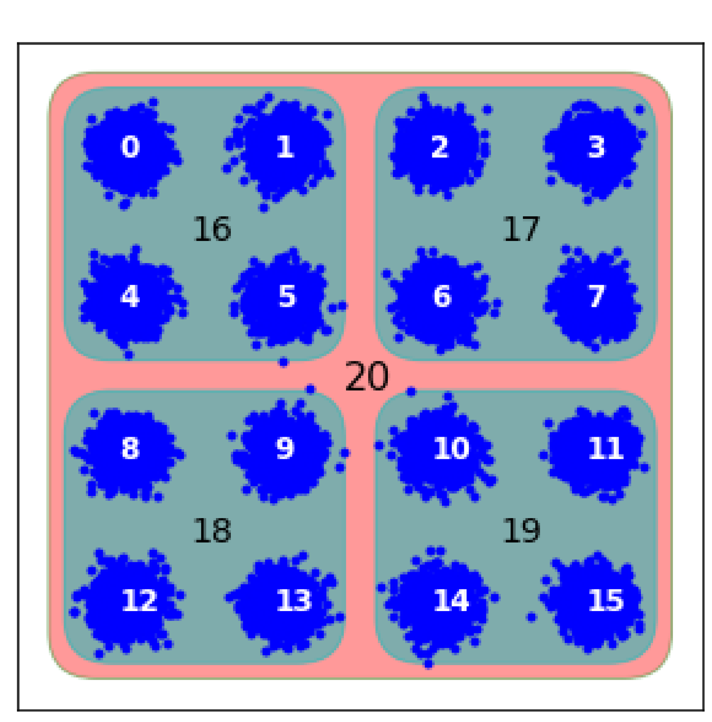

Figure: Gaussian used for the synthetic experiment

- 16 gaussians corresponds to a single label $l_1, l_2 \ldots l_{16}$.
- 3 layered tree hierarchy of labels

Prob	0.00		0.20		0.40	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
$\text{HiddeN}_{\text{flt}}$	96.8	89.1	93.2	87.8	90.4	87.7
$\text{HiddeN}_{\text{cas}}$	98.0	93.4	94.4	88.9	91.9	91.0
$\text{HiddeN}_{\text{jnt}}$	98.1	94.0	94.8	91.6	92.3	91.7

Table: Synthetic data here has 12000 training and 8000 test samples.

Experiments

Dataset	Hierarchy	Hyperbolicity	|L|	Avg(|L|)	Max(|L|)	Train	Val	Test
RCV1	Tree	0	104	3.24	17	20833	2314	781265
NYT	Tree	1	120	6.58	24	86461	9606	9903
Yelp	DAG	1	539	4.07	32	98460	10939	46884

Table: Statistics of the datasets.

Dataset	Method	Micro-F1	Macro-F1
RCV1	TextCNN-Flat*	76.6	43.0
	$\text{HiddeN}_{\text{flt}}$	77.9	44.5
	$\text{HiddeN}_{\text{cas}}$	78.0	45.5
	$\text{HiddeN}_{\text{jnt}}$	**79.3**	**47.3**
NYTimes	TextCNN-Flat*	69.5	39.5
	$\text{HiddeN}_{\text{flt}}$	76.4	37.1
	$\text{HiddeN}_{\text{cas}}$	74.6	33.2
	$\text{HiddeN}_{\text{jnt}}$	**77.0**	**43.6**
Yelp	TextCNN-Flat*	**62.8**	27.3
	$\text{HiddeN}_{\text{flt}}$	62.5	**37.9**
	$\text{HiddeN}_{\text{cas}}$	60.5	33.9
	$\text{HiddeN}_{\text{jnt}}$	60.8	35.6

Table: Performance comparison on all three datasets with TextCNN as the base classification model.

Dataset	Method	Micro-F1	Macro-F1
RCV1	$\text{HiddeN}_{\text{euc}}$	78.4	47.6
	$\text{HiddeN}_{\text{jnt}}$	**79.3**	**47.3**
NYTimes	$\text{HiddeN}_{\text{euc}}$	76.4	40.4
	$\text{HiddeN}_{\text{jnt}}$	**77.0**	**43.6**
Yelp	$\text{HiddeN}_{\text{euc}}$	**61.1**	34.2
	$\text{HiddeN}_{\text{jnt}}$	60.8	**35.6**

Table: Performance comparison for $\text{HiddeN}_{\text{jnt}}$ with $\text{HiddeN}_{\text{euc}}$.

Dataset	$\text{HiddeN}_{\text{jnt}}$		HiLAP	
	Micro	Macro	Micro	Macro
RCV1	**79.3**	47.3	78.6	**50.5**
NYTimes	**77.0**	**43.6**	69.9	43.2
Yelp	60.8	35.6	**65.5**	**37.3**

Table: Performance comparison of $\text{HiddeN}_{\text{jnt}}$ with HiLAP

	$\text{HiddeN}_{\text{flt}}$	$\text{HiddeN}_{\text{jnt}}$	$\text{HiddeN}_{\text{cas}}$
RCV1	21.2	**53.9**	44.1
NYTimes	11.4	**39.5**	36.1
Yelp	16.3	**31.9**	28.8

Table: Spearman rank correlation test for the generated embeddings for all the datasets. Each method is compared against the ground truth hierarchy.

References

[1] Yuning Mao and J et al. Tian.
Hierarchical text classification with reinforced label assignment (hilap).
In *Proceedings of the 2019 Conference of EMNLP-IJCNLP*, pages 445–455, 2019.

Acknowledgements

We thank Bamdev Mishra and Pratik Jawanpuria (Microsoft India, Hyderabad) for valuable discussions that gave us impetus to work towards this problem. Ayush Maheshwari is supported by a Fellowship from Ekal Foundation (www.ekal.org).