Poincaré GloVe: Hyperbolic Word Embeddings

Alexandru Țifrea, Gary Bécigneul, Octavian-Eugen Ganea
Department of Computer Science, ETH Zürich, Switzerland

Problem Definition

Goal 1: train unsupervised word embeddings that capture word similarity and word analogy, like GloVe

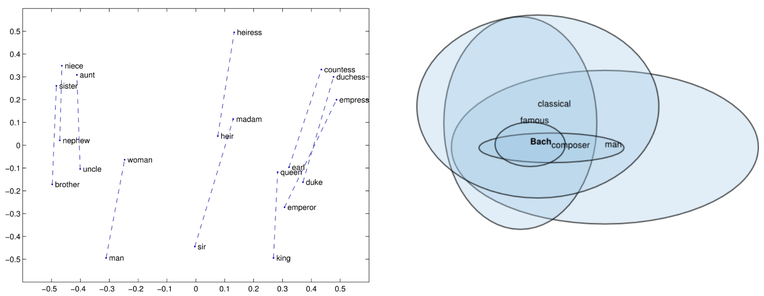
Goal 2: additionally capture lexical entailment

- use the hyperbolic space for training
- no supervision during training
- provide mathematically sound way to perform inference for the three evaluation tasks

Prior work & Baselines

1) **GloVe embeddings** (Pennington et al, 2014)

- Euclidean point embeddings
- capture word similarity and analogy



2) **Gaussian Word Embeddings** (Vilnis et al, 2014)

- represent words in the space of diagonal Gaussians
- capture lexical entailment, but no indication of their performance on word analogy tasks

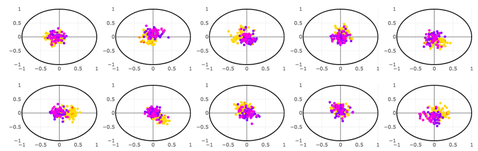
Poincaré GloVe

Loss: $J = \sum_{i,j=1}^{V} f(X_{ij}) \Big(-h(\mathrm{d}_{\text{Poincaré}}(w_i, \tilde{w}_j)) + b_i + \tilde{b}_j - \log X_{ij} \Big)^2$

- $\mathrm{d}_{\text{prod}}(x, y) = \sqrt{\sum_{i=0}^{k} \mathrm{d}_{\text{Poincaré}}(x_i, y_i)^2}$ for training in the Cartesian product of Poincaré balls

Connection to word2gauss

Let $\Sigma = diag(\sigma)^2$, $\Sigma' = diag(\sigma')^2$ and $d_{\mathbb{H}^2}(\cdot,\cdot)$ be the half-plane distance. Then $d_F\left(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\mu',\Sigma')\right) = \sqrt{\sum_{i=1}^{n} 2 d_{\mathbb{H}^2}\left((\mu_i/\sqrt{2}, \sigma_i), (\mu_i'/\sqrt{2}, \sigma_i')\right)^2}$ (Costa et al, 2015).



Word analogy via Parallel Transport

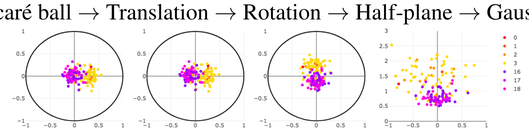
"*king* is to *queen* what *man* is to *woman*"

$$A \to B \Rightarrow C \to D, \quad \text{and} \quad A \to C \Rightarrow B \to D$$
$$d_1 = c \oplus gyr[c, \ominus a](\ominus a \oplus b), \quad \text{and} \quad d_2 = b \oplus gyr[b, \ominus a](\ominus a \oplus c)$$

Solution: interpolate between the two points obtained with parallel transport

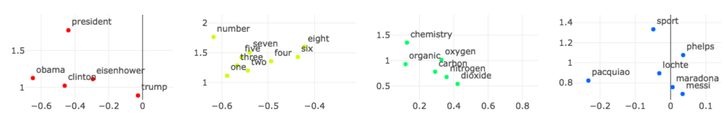
Lexical entailment

Poincaré ball → Translation → Rotation → Half-plane → Gaussians



Translation & rotation parameters chosen using two approaches:
1) semi-supervised (using a sample of WordNet generic/specific words)
2) unsupervised (using sample of frequent/rare words)

Let P, Q be two words with Gaussian embeddings (μ, Σ) and (μ', Σ').
LE score: is-a$(P, Q) = \log(V_{\Sigma'}) - \log(V_{\Sigma}) = \sum_{i=1}^{n} (\log(\sigma_i') - \log(\sigma_i))$
Intuition:
Large variance ⇒ Generic word
Small variance ⇒ Specific word



Experiments

- Word similarity and analogy results (highlighted: the **best** and the 2nd best).

Experiment name	RareWord	WordSim	SimLex	SimVerb	Google	MSR
100D Vanilla GloVe	0.3798	0.5901	0.2963	0.1675	0.5931	0.4868
100D Vanilla GloVe w/ init trick	0.3787	0.5668	0.2964	0.1639	0.6167	0.4826
100D Poincaré GloVe $h(x) = \cosh^2(x)$, w/ init trick	0.4187	0.6209	**0.3208**	**0.1915**	**0.6339**	**0.4971**
50x2D Poincaré GloVe $h(x) = \cosh^2(x)$, w/ init trick	**0.4276**	**0.6234**	0.3181	0.189	0.6045	0.4849
50x2D Poincaré GloVe $h(x) = x^2$, w/ init trick	0.4104	0.5782	0.3022	0.1685	0.6300	0.4672

- Hyperlex results (Spearman correlation) for different model types ordered according to their difficulty.

Table 1

Model type	Method	ρ
Supervised embedding learning & **Unsupervised** hypernymy score	WN-Eucl from Nickel et al	0.389
	WN-Poincaré from Nickel et al	**0.512**
Unsupervised embedding learning & **Weakly-supervised** hypernymy score	50x2D Poincaré GloVe, $h(x) = \cosh^2(x)$, init trick (190k)	
	• WordNet 20+20	0.360
	• WordNet 400+400	0.402
	50x2D Poincaré GloVe, $h(x) = x^2$, init trick (190k)	
	• WordNet 20+20	0.344
	• WordNet 400+400	**0.421**
Unsupervised embedding learning & **Unsupervised** hypernymy score	Word2Gauss-DistPos	0.206
	SBOW-PPMI-CΔS from Chang et al	**0.345**
	50x2D Poincaré GloVe, $h(x) = \cosh^2(x)$, init trick (190k) • Unsupervised 5k+5k	0.284
	50x2D Poincaré GloVe, $h(x) = x^2$, init trick (190k) • Unsupervised 5k+5k	**0.341**

Resources

Code: https://github.com/alex-tifrea/poincare_glove
Blog post: http://hyperbolicdeeplearning.com/poincare-glove/

References

- Pennington et al, "Glove: Global vectors for word representation", EMNLP 2014
- Vilnis et al, "Word representations via Gaussian embedding", ICLR 2015
- Costa et al, " Fisher information distance: a geometrical reading", 2015
- Chang et al, "Distributional inclusion vector embedding for unsupervised hypernymy detection", NAACL 2018
- Nickel et al, "Poincaré embeddings for learning hierarchical representations", NIPS 2017.