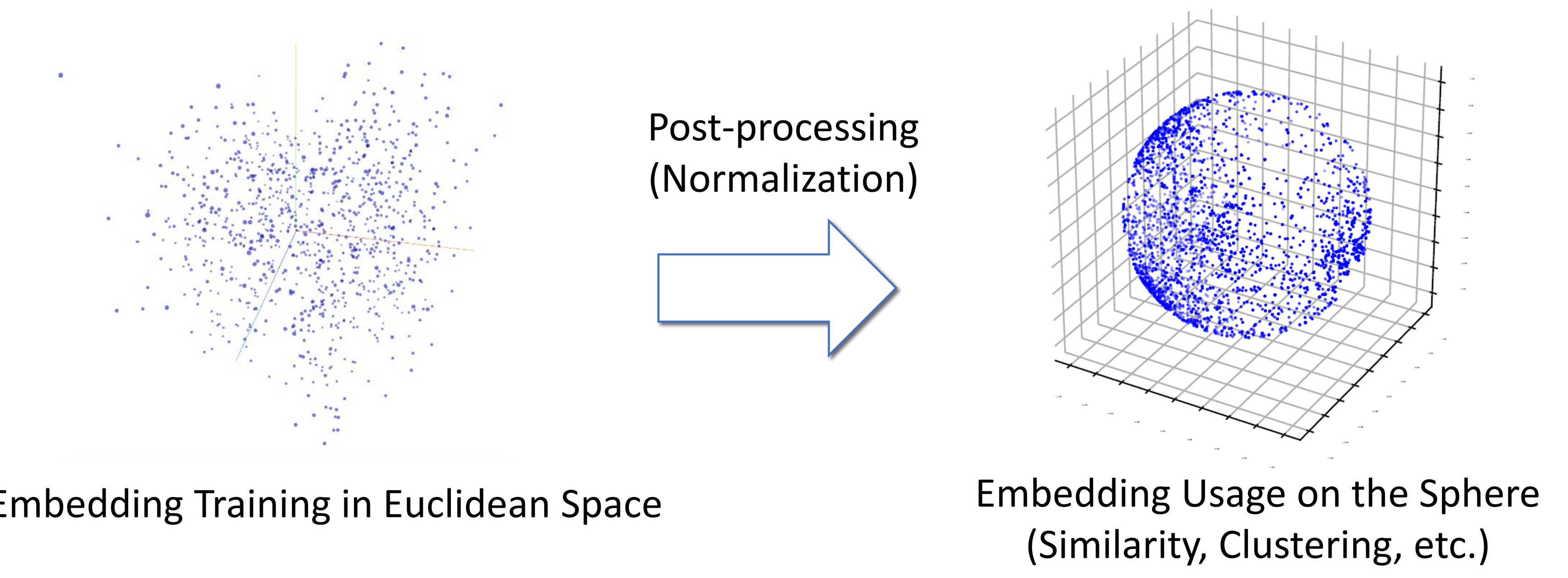
Spherical Text Embedding

Yu Meng[1], Jiaxin Huang[1], Guangyuan Wang[1], Chao Zhang[2], Honglei Zhuang[3], Lance Kaplan[4] and Jiawei Han[1]
University of Illinois at Urbana-Champaign[1], Georgia Institute of Technology[2], Google Research[3], U.S. Army Research Laboratory[4]

Paper: Code:

Introduction

- Text Embedding is a milestone in NLP and ML
- **Directional (cosine) similarity** is more effective for embedding applications



- The objective optimized is not really the one we use

Embedding dot product is optimized

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^\top v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^\top v_{w_I}\right)}$$

Word2Vec

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

GloVe

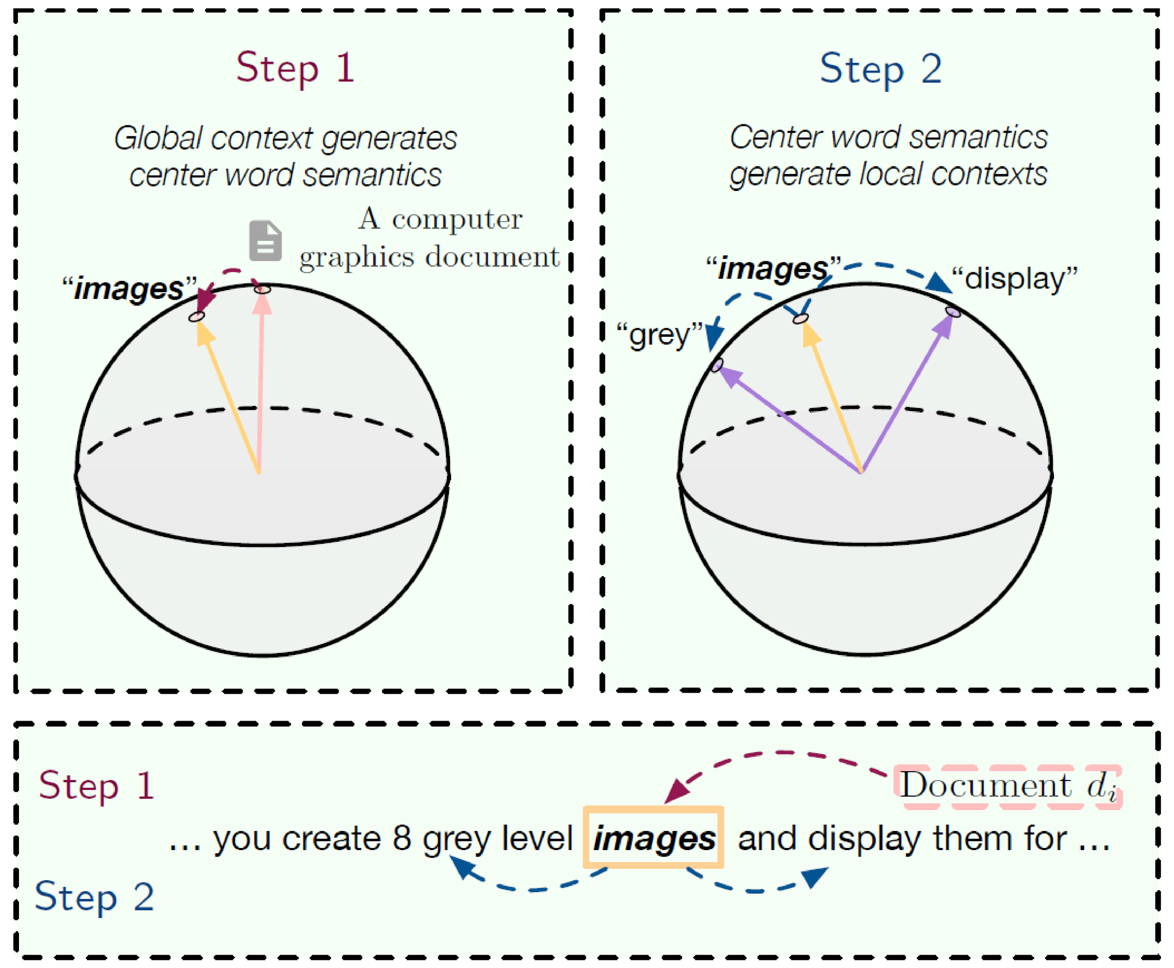
- Inconsistency between training and usage

	Metrics	A: *lover-quarrel*		B: *rock-jazz*	
Training	Dot Product	5.284	<	6.287	Inconsistency
Usage	Cosine Similarity	0.637	>	0.628	

- Spherical Text Embedding
  - Train embeddings on the unit sphere
  - Jointly learn word and document/paragraph embeddings
  - State-of-the-art on various embedding applications

Model & Optimization

- Spherical Generative Model (two-step generation):



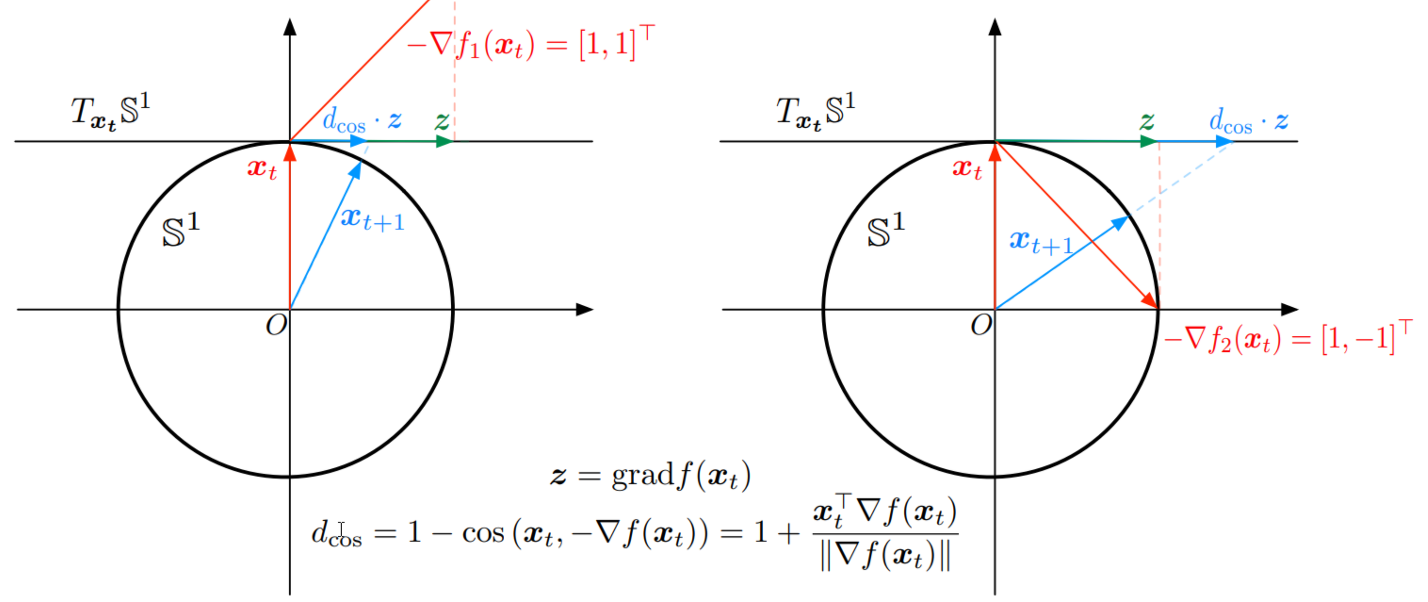
- The generative probability is characterized by vMF distribution (Theorem 1)
- Objective: $\mathcal{L}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{d}) = \max\Big(0, m - \log\left(c_p(1)\exp(\cos(\boldsymbol{v},\boldsymbol{u}))\cdot c_p(1)\exp(\cos(\boldsymbol{u},\boldsymbol{d}))\right)$

$$+ \log\left(c_p(1)\exp(\cos(\boldsymbol{v},\boldsymbol{u}'))\cdot c_p(1)\exp(\cos(\boldsymbol{u}',\boldsymbol{d}))\right)\Big)$$

$$= \max\left(0, m - \cos(\boldsymbol{v},\boldsymbol{u}) - \cos(\boldsymbol{u},\boldsymbol{d}) + \cos(\boldsymbol{v},\boldsymbol{u}') + \cos(\boldsymbol{u}',\boldsymbol{d})\right).$$

- Riemannian optimization with angular distance:

$$\boldsymbol{x}_{t+1} = R_{\boldsymbol{x}_t}\left(-\eta_t\left(1 + \frac{\boldsymbol{x}_t^\top \nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}\right)\left(I - \boldsymbol{x}_t\boldsymbol{x}_t^\top\right)\nabla f(\boldsymbol{x}_t)\right)$$



Evaluations

- Word Similarity:

Table 1: Spearman rank correlation on word similarity evaluation.

Embedding Space	Model	WordSim353	MEN	SimLex999
Euclidean	Word2Vec	0.711	0.726	0.311
	GloVe	0.598	0.690	0.321
	fastText	0.697	0.722	0.303
	BERT	0.477	0.594	0.287
Poincaré	Poincaré GloVe	0.623	0.652	0.321
Spherical	**JoSE**	**0.739**	**0.748**	**0.339**

- Document Clustering:

Table 2: Document clustering evaluation on the 20 Newsgroup dataset.

Embedding	Clus. Alg.	MI	NMI	ARI	Purity
Avg. W2V	K-Means	1.299 ± 0.031	0.445 ± 0.009	0.247 ± 0.008	0.408 ± 0.014
	SK-Means	1.328 ± 0.024	0.453 ± 0.009	0.250 ± 0.008	0.419 ± 0.012
SIF	K-Means	0.893 ± 0.028	0.308 ± 0.009	0.137 ± 0.006	0.285 ± 0.011
	SK-Means	0.958 ± 0.012	0.322 ± 0.004	0.164 ± 0.004	0.331 ± 0.005
BERT	K-Means	0.719 ± 0.013	0.248 ± 0.004	0.100 ± 0.003	0.233 ± 0.005
	SK-Means	0.854 ± 0.022	0.289 ± 0.008	0.127 ± 0.003	0.281 ± 0.010
Doc2Vec	K-Means	1.856 ± 0.020	0.626 ± 0.006	0.469 ± 0.015	0.640 ± 0.016
	SK-Means	1.876 ± 0.020	0.630 ± 0.007	0.494 ± 0.012	0.648 ± 0.017
JoSE	K-Means	1.975 ± 0.026	0.663 ± 0.008	0.556 ± 0.018	0.711 ± 0.020
	SK-Means	**1.982** ± 0.034	**0.664** ± 0.010	**0.568** ± 0.020	**0.721** ± 0.029

- Document Classification:

Table 3: Document classification evaluation using k-NN ($k = 3$).

Embedding	20 Newsgroup Macro-F1	20 Newsgroup Micro-F1	Movie Review Macro-F1	Movie Review Micro-F1
Avg. W2V	0.630	0.631	0.712	0.713
SIF	0.552	0.549	0.650	0.656
BERT	0.380	0.371	0.664	0.665
Doc2Vec	0.648	0.645	0.674	0.678
JoSE	**0.703**	**0.707**	**0.764**	**0.765**

- Training Efficiency:

Table 4: Training time (per iteration) on the latest Wikipedia dump.

Word2Vec	GloVe	fastText	BERT	Poincaré GloVe	JoSE
0.81 hrs	0.85 hrs	2.11 hrs	> 5 days	1.25 hrs	**0.73 hrs**