Learning Word Embedding Distributions

B.Shreya cs15btech11009
Joint work with Jayashree P, Dr.Srijith P.K

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Timeline

Jan — Read papers, understood KL approximation, implemented energy function based on KL divergence

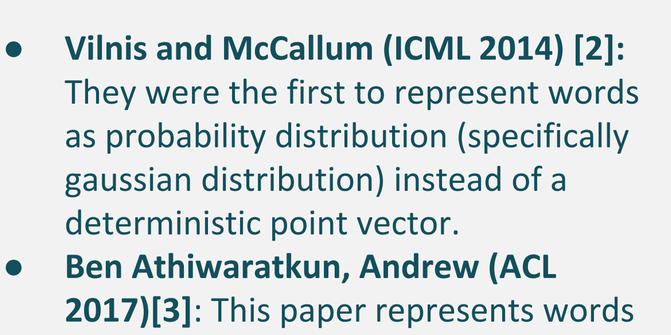
Feb — Implemented energy function based on KL divergence ,Implemented model which share same means,variances across all mixtures

Mar — Implemented clustering based model where means and variances are shared within clusters

April — Implemented various metrics to compare old approaches and current approach

Introduction

- Learning representations for words is of high importance lately.
- Words can be represented using vector space models as embeddings.
- Whereas, representing words as gaussian distribution helps in capturing the uncertainty information.
- But having only one representation per word doesn't really help in capturing meanings of polysemous words.



Literature

There have been several approaches for word distributions.

- **Tomas Mikolov, Kai Chen(2013) [1]:** This is popularly known as word2vec and it uses continuous bag of words and skip-gram models for generating vector representations.
- **Vilnis and McCallum (ICML 2014) [2]:** They were the first to represent words as probability distribution (specifically gaussian distribution) instead of a deterministic point vector.
- **Ben Athiwaratkun, Andrew (ACL 2017)[3]**: This paper represents words as Gaussian mixtures to capture multiple sense words using energy based max-margin objective.

Word Representation

- Each word w is represented as:

$$f_w(\vec{x}) = \sum_{i=1}^{K} p_{w,i}\, \mathcal{N}\left[\vec{x}; \vec{\mu}_{w,i}, \Sigma_{w,i}\right]$$

- Here $\mu_{w,i}$ represents mean vector of ith component of w, $\Sigma_{w,i}$ is the component covariance matrix and $p_{w,i}$ represents the component probability.

Methodology

- Expected likelihood kernel is used as an energy measure in the loss function.
- KL divergence is a better measure for word embeddings than expected likelihood as it is an asymmetric and can capture entailment information.
- The problem is that KL divergence doesn't have a closed form if two distributions are gaussian mixtures.
- We considered the model from MultiModal Word distributions paper.

Approximation for KL

An approximation of the KL divergence algorithm is implemented for a mixture of gaussians. [4]

$$\underbrace{\sum_a \omega_a^f \log \frac{\sum_\alpha \omega_\alpha^f e^{-D_{\mathrm{KL}}(f_a||f_\alpha)}}{\sum_b \omega_b^g t_{ab}} - \sum_a \omega_a^f H(f_a)}_{D_{\mathrm{lower}}(f||g)}$$

$$\underbrace{\sum_a \omega_a^f \log \frac{\sum_\alpha \omega_\alpha^f z_{a\alpha}}{\sum_b \omega_b^g e^{-D_{\mathrm{KL}}(f_a||g_b)}} + \sum_a \omega_a^f H(f_a)}_{D_{\mathrm{upper}}(f||g)}$$

$$t_{ab} \triangleq \int_{\mathbf{x}} f_a(\mathbf{x}) g_b(\mathbf{x}) d\mathbf{x} \qquad z_{a\alpha} \triangleq \int_{\mathbf{x}} f_a(\mathbf{x}) f_\alpha(\mathbf{x}) d\mathbf{x}$$

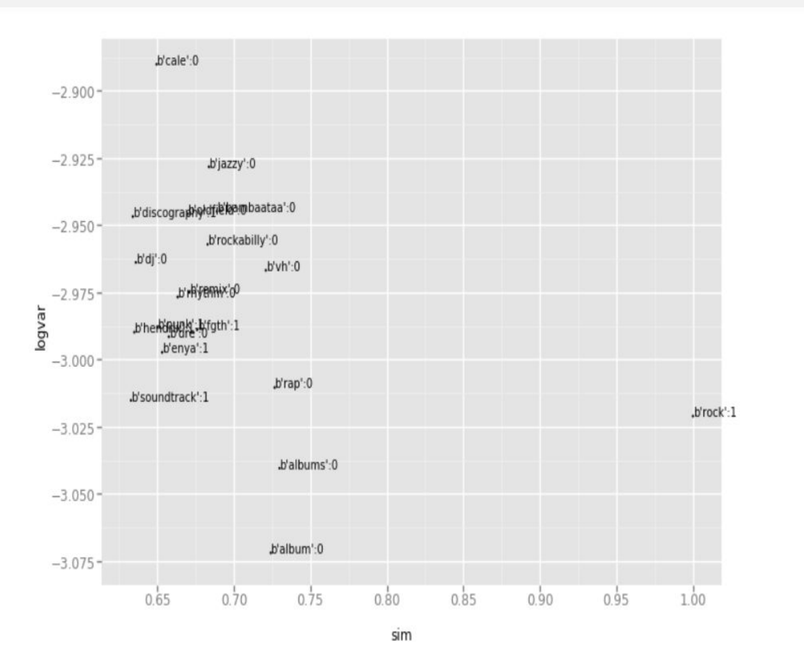
f_a represents gaussian mixture corresponding to component a of word1, g_b represents gaussian mixture corresponding to component b of word2. D_{KL} is KL divergence between gaussians, H(f) represents entropy, D_{lower} and D_{upper} are the lower and upper bounds of approx KL divergence.

$$D_{\mathrm{mean}}(f||g) \triangleq [D_{\mathrm{upper}}(f||g) + D_{\mathrm{lower}}(f||g)]/2$$

Results

Nearest neighbors for our approach for the word 'rock':
Component 0 : thin, sedimentary, molten, granite,felsic
Component 1: albums, rap, vh, bambaata, jazzy

Nearest neighbors for multimodal approach for the word 'rock':
Component 0 : platinum, metal, agate, bubbling, chamois
Component 1: roll, funk, blues, bands, floyd



Graph for nearest neighbors of one of the components of 'rock' using dot product

Model	Spearman Correlation using maxdot	Spearman Correlation using max(neg KL divergence)
Multimodal (w2gm)	43.37	30.20
Our approach	44.41	31.22

Spearman correlation on SCWS dataset for **maxdot** and **KL divergence** metrics

Conclusion

- Using mixture of gaussians for representing a word helps in capturing multiple senses.
- The results using KL divergence as a energy metric in Loss function proved to be better than EL kernel.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean 2013a, Efficient estimation of word representations in vector space

[2] Luke Vilnis and Andrew McCallum 2014, Word representations via gaussian embedding

[3] Ben Athiwaratkun, Andrew Gordon Wilson, 2017, Multimodal Word Distributions

[4] J.-L. Durrieu,J.-Ph. Thiran, F. Kelly 2012, Lower and Upper Bounds For Approximation of The KL Divergence Between Gaussian Mixture Models