SUMMARY

Multimodal Word Distributions ACL 2017

Introduction:

This paper introduces multimodal word embeddings where each word is represented as a mixture of gaussian mixtures and each gaussian refers to a different meaning of the word in case of polysemy words.

Previous Approaches:

Word2vec - In this approach,each word is represented as a single vector where words with similar meanings are nearer to one another.
Single Gaussian Representation - In this approach, each word is represented as a gaussian distribution where mean and covariance are learned from the data. This provides much richer representation than point vectors but it cannot handle polysemies.

To overcome the problem in the above approaches, this paper proposes represent each word as a mixture of gaussians where each gaussia corresponds to different meanings in case of polysemies.

Approach:

Each word w in a dictionary is represented as a gaussian mixture with k components and the distribution of the word is given by:

The main goal is to learn the model parameters from a corpus of natural sentences.

Each mean vector of the gaussian ca represent one meaning of a word. They use a maximum margin energy-based ranking objective function which is given by:

$$\begin{aligned} L_\theta(w, c, c') = \max(0, & \\ & m - \log E_\theta(w, c) + \log E_\theta(w, c')) \end{aligned}$$

and the energy function is given by -

$$E(f, g) = \int f(x) g(x)\, dx = \langle f, g \rangle_{L_2}$$

$$\log E_\theta(f, g) = \log \sum_{j=1}^{K} \sum_{i=1}^{K} p_i q_j e^{\xi_{i,j}}$$

$$\begin{aligned} \xi_{i,j} \equiv & \log \mathcal{N}(0; \vec{\mu}_{f,i} - \vec{\mu}_{g,j}, \Sigma_{f,i} + \Sigma_{g,j}) \\ & = -\frac{1}{2} \log \det(\Sigma_{f,i} + \Sigma_{g,j}) - \frac{D}{2} \log(2\pi) \\ -\frac{1}{2} & (\vec{\mu}_{f,i} - \vec{\mu}_{g,j})^\top (\Sigma_{f,i} + \Sigma_{g,j})^{-1} (\vec{\mu}_{f,i} - \vec{\mu}_{g,j}) \end{aligned}$$

Here c is nearby word to w within a context window of l, and c' is a negative context word.

It is based on the concept that the words nearer to one another are similar. This loss function tries to push the similarity of a word and its positive context higher than the similarity of a word and its negative context by a margin of m.

The energy function captures similarity between ith component of word wf$_f$and jth component of word w$_g$. So higher the similarity between these values, higher the energy value and lower will be the loss function.

Experiments:

They use a concatenation of 2 datasets, UKWAC (2.5 billion tokens) and Wackypedia (1 billion tokens). For each gaussian mixture component of a word the mean represents the embedding of the ith component and the variance represents its uncertainty. D=50, K=2 and l=10 for the experiments.

Since each word has multiple components the similarity scores that are used between two words are:

1. Expected Likelihood Kernel - Inner product between gaussian mixtures
2. Maximum Cosine Similarity - The maximum of cosine similarity between all pairs of components of two words w_f and w_g.
3. Minimum Euclidean Distance - The minimum of euclidean distance between all pairs of components of two words w_f and w_g.

The word embeddings are evaluated on several word similarity datasets and spearman correlation is calculated between the labels and the scores. The model w2gm outperforms all the other state-of-art models using various similarity measures.

Using a mixture of gaussians succeeds in modeling word uncertainty better than unimodal approaches as it has reduced variances for all components rather than high variance for a single gaussian in case of polysemous words.

Conclusion:

The multimodal word representation introduced in this paper is successful in capturing different semantics of polysemous words, uncertainty, and entailment, and also perform favorably on word similarity benchmarks.

Probabilistic FastText for Multi-Sense Word Embeddings
ACL 2018

Introduction:

This paper introduces word embeddings with multiple gaussian mixtures where mean of each mixture component is given by sum of n-grams. This model can handle not only words with multiple meanings but also rare misspelt or even unseen words.

Previous Approaches:

Most of the word embeddings are based on dictionary-level embeddings and cannot handle rare/unseen words. To overcome this problem,FASTTEXT was introduced which used character-level embeddings where each word is modeled as a sum of n-gram vectors.

Approach:

Each word is represented as a gaussian mixture.
Each component of the mixture can represent different word senses, and the mean vectors of each component decompose into vectors of n-grams, to capture character-level information. This embedding has the ability to handle rare words and even foreign polysemies with an improvement of 1% over the w2gm model on SCWS.
The overall approach is similar to that of w2gm except to the mean vector calculation in the gaussian mixture. The mean vectors in w2gm are dictionary level which leads to poor semantic estimate for rare words.To overcome this problem, using subword structures we can have the mean as :

$$\mu_w = \frac{1}{|NG_w| + 1}\left(v_w + \sum_{g \in NG_w} z_g\right)$$

Experiments:

The probabilistic FASTTEXT which combines subword structure with probabilistic embeddings is trained on UKWAC and Wackypedia datasets for English and for foreign languages it is trained on French, German, and Italian text corpuses. Parameters are fixed at K=2,l=10 and n=3,4,5.

It is observed that the subword embeddings prefer words with overlapping characters as nearest neighbors. This model is seen to outperform all state of art models including w2gm by 3.1% and FASTTEXT by 1.2%. Even in case of foreign languages, this model outperforms the others.

Conclusion:

This paper thus proposes the first ever model to handle polysemies rare and unseen words.

Future work lies in:

1. Trade-off between learning full covariance matrices for each word distribution, computational complexity, and performance as in this model we are directly using spherical covariance matrices.
2. Co-training on many languages.