# Bayesian Paragraph Vectors

Geng Ji[1], Robert Bamler[2], Erik B. Sudderth[1] and Stephan Mandt[2]
[1]University of California, Irvine; [2]Disney Research

## Introduction

- Many tasks in natural language processing require fixed-length features for text passages of variable length, such as sentences, paragraphs, or documents.
- We propose an unsupervised generative model whose maximum likelihood solution recovers the traditionally neural-network based *paragraph vectors* [1].
- This probabilistic formulation allows us to go beyond point estimates of parameters and to perform Bayesian posterior inference.
- We find that the entropy of paragraph vectors decreases with the length of documents, and that information about posterior uncertainty improves performance in downstream supervised learning tasks.

## Generative Process

We propose a probabilistic model that reinterprets the doc2vec approach [1]. We achieve this by extending the recent Bayesian interpretation [2] of word2vec [3].

- For each word $i$ in the vocabulary, sample its word embedding vector $U_i$ and context embedding vector $V_i$ from a Gaussian prior:

$$U_i \sim \mathcal{N}(0, \lambda^2 I), \quad V_i \sim \mathcal{N}(0, \lambda^2 I)$$

- For each document $n$, sample its paragraph vector $d_n$ from another Gaussian:

$$d_n \sim \mathcal{N}(0, \phi^2 I)$$

- Draw each pair of words $(i, j)$ for document $n$ uniformly from the vocabulary, and assign it with a binary label $z_{n,ij}$:

$$z_{n,ij} \sim \mathrm{Bern}\big(\sigma(U_i^\top (V_j + d_n))\big), \quad \sigma(x) \triangleq 1/(1+e^{-x})$$

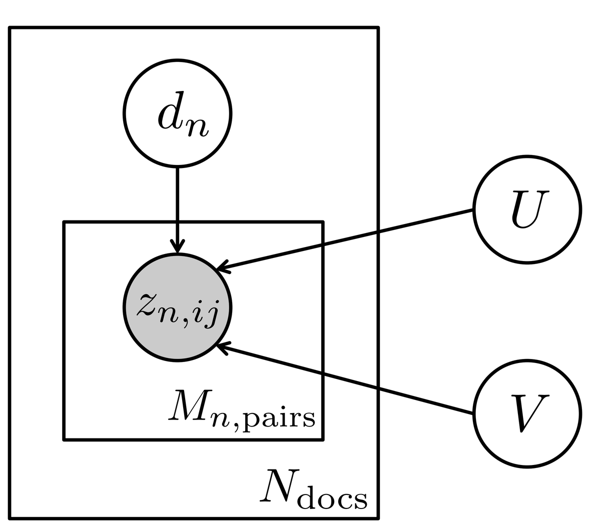


**Figure 1.** Bayesian paragraph vectors

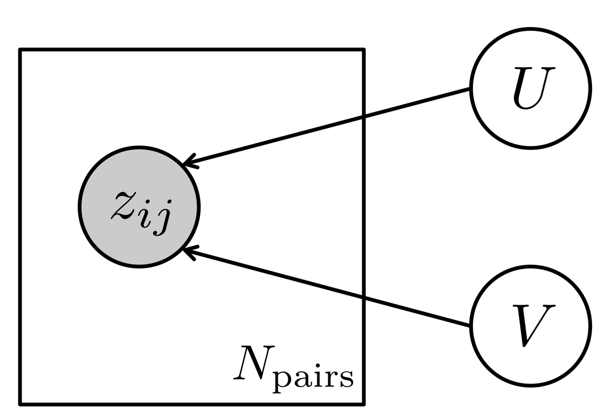


**Figure 2.** Bayesian skip-gram model [2]

- Positive labels ($z_{n,ij} = 1$) correspond to occurrences of the word $i$ in the context of word $j$ somewhere in document $n$; negative examples ($z_{n,ij} = 0$) are artificial evidence constructed by sampling from the noise distribution $P(i,j) \propto f(i) f(j)^{\frac{3}{4}}$ as in [3], where $f$ is the empirical unigram frequency across the training corpus.

$$p(z_{n,ij} \mid U_i, V_j, d_n) = \sigma\big(U_i^\top (V_j + d_n)\big)^{z_{n,ij}} \, \sigma\big(-U_i^\top (V_j + d_n)\big)^{1 - z_{n,ij}}$$

## Edward Implementation

Bayesian paragraph vectors can be easily specified in Edward, a Python library for probabilistic modeling and inference [4]:

```
from edward.models import Bernoulli, Normal
U = Normal(loc=tf.zeros((W, E), dtype=tf.float32), scale=lam)
V = Normal(loc=tf.zeros((W, E), dtype=tf.float32), scale=lam)
d_n = Normal(loc=tf.zeros(E, dtype=tf.float32), scale=phi)
u_n = tf.nn.embedding_lookup(U, indices_n_I)
v_n = tf.nn.embedding_lookup(V, indices_n_J)
z_n = Bernoulli(logits=tf.reduce_sum(u_n * (v_n + d_n), axis=1))
```

## Black-box Variational Inference

**Idea**: We expect a broader posterior distribution for the local paragraph vectors. Thus we use MAP inference for the global word embeddings $U$ and $V$, and fit a variational Gaussian distribution (fully factorized) for the local variable $d_n$ of each document.

- **Stage 1**: train the global word embeddings via stochastic gradient descent.
  0. Each mini-batch contains a document and a fixed set of negative examples.
  1. Update its paragraph vector until convergence. As this local optimization is noise free, it converges quickly under a constant learning rate.
  2. Perform a single gradient step for the global variables. This gradient is noisy due to the mini-batch sampling and the stochastic generation of negative examples. For this reason, a decreasing learning rate is used.
  3. Reinitialize the paragraph vector and proceed to the next document.
- **Stage 2**: fit a variational Gaussian for each paragraph vector using BBVI [5].
  - Hold fixed the global variables $U$ and $V$ learned in stage 1.
  - Use BBVI with re-parameterization gradients [6,7] provided in Edward.
  - Generate new negative examples in each update step to avoid overfitting.
  - Use a decreasing learning rate to perform the stochastic optimization above.

## Experiments

- We use Bayesian paragraph vectors as input features for two binary classification tasks in natural language processing: sentiment analysis and paraphrase detection.
- **Finding 1**: the posterior uncertainty of Bayesian paragraph vectors decreases as the length of paragraphs grows:

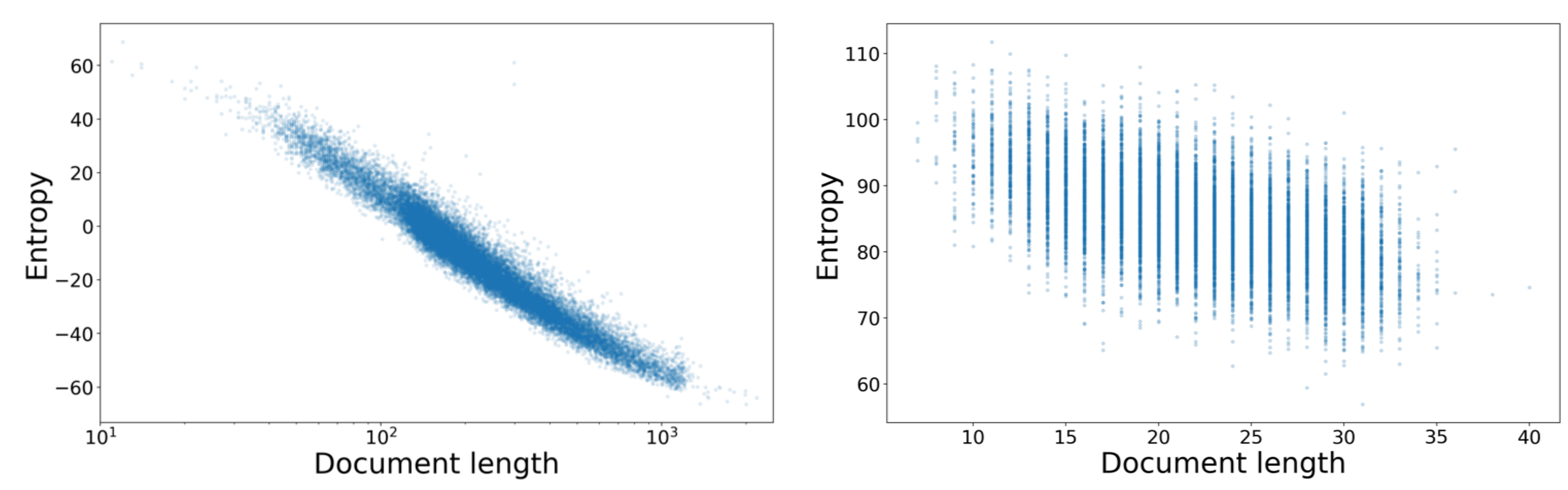


**Figure 3.** Entropy of paragraph vectors as a function of the number of words in each document. Left: movie reviews in the IMDB dataset. Right: news clips in the MSR dataset.

- **Finding 2**: by concatenating the variational mean and standard deviation features inferred by BBVI, we improve the classification accuracy compared to MAP.

**Table 1.** Classification accuracy of MAP and black-box variational inference.

| Task (dataset) | MAP | BBVI |
|---|---|---|
| Sentiment Analysis (IMDB) | 86.9 | **87.0** |
| Paraphrase detection (MSR) | 70.0 | **71.0** |

## Discussion

- Our experiments verify that paragraph vectors of short documents have higher uncertainty, and that knowledge of it improves downstream supervised tasks.
- In addition to MAP and VI, we experimented with Hamiltonian Monte Carlo (HMC) inference, but our preliminary results showed worse performance. A possible reason is that we had to use a fixed set of negative examples for each doc when generating HMC samples, which may result in overfitting to the noise.
- We believe that more sophisticated models of document embeddings would also benefit from a Bayesian treatment of the local variables.

## Contact

Geng Ji
Computer Science Department
UC Irvine
Email: gji1@uci.edu
Phone: +1 (949) 701-7920

## References

1. Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML*.
2. Barkan, O. (2017). Bayesian neural word embedding. In *AAAI*.
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*.
4. Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv:1610.09787*.
5. Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *AISTATS*.
6. Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR*.
7. Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.