Variational Autoencoders for Topic Modeling

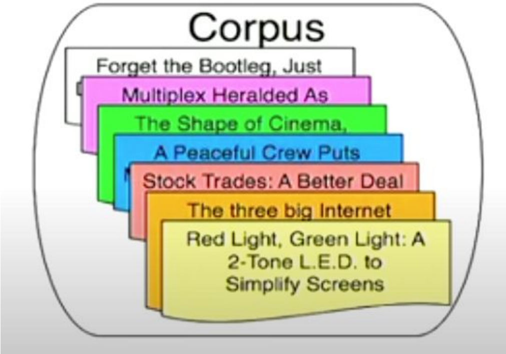
Kelly Geyer

April 29, 2020

Task

- We want to improve upon topic modeling using deep learning.
- Challenge: How to evaluate the relative performance of topic models?

Given: A collection of documents, number of topics K



"Topic modeling"

Black Box Algorithm

Returns: (i) Cluster of words by topic

"technology" "business" "film"

Returns: (ii) Cluster of document by topic

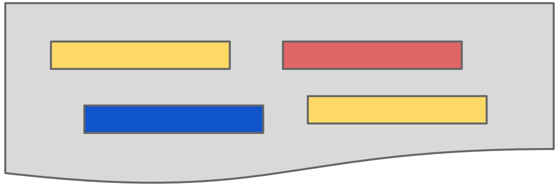


Image credit: https://www.youtube.com/watch?v=yK7nN3FcgUs

Motivation: Latent Dirichlet Analysis

Latent Dirichlet Analysis (LDA) generates words and topics from a generative process:

for *each document* **w do**
  Draw topic distribution $\theta \sim \text{Dirichlet}(\alpha)$;
  for *each word at position* n **do**
    Sample topic $z_n \sim \text{Multinomial}(1, \theta)$;
    Sample word $w_n \sim \text{Multinomial}(1, \beta_{z_n})$;
  end
end

Use Bayes' Rule to estimate posteriors for:

θ (topic-doc scores) and

β_{z_n} (topic-doc-word scores)

Drawbacks

- Optimization formulaizations of LDA are inflexible for new model distributions
  - e.g., lack of closed form solutions)
- The distribution of p(wltheta,beta) is a mixture of multinomial distributions
  - Often results in poor quality topics that do not correspond well with human judgment

Related work: ProdLDA

LDA

for *each document* **w do**
  Draw topic distribution $\theta \sim \text{Dirichlet}(\alpha)$;
  for *each word at position* n **do**
    Sample topic $z_n \sim \text{Multinomial}(1, \theta)$;
    $\boxed{\text{Sample word } w_n \sim \text{Multinomial}(1, \beta_{z_n});}$
  end
end

ProdLDA

β_{z_n} is unnormalized

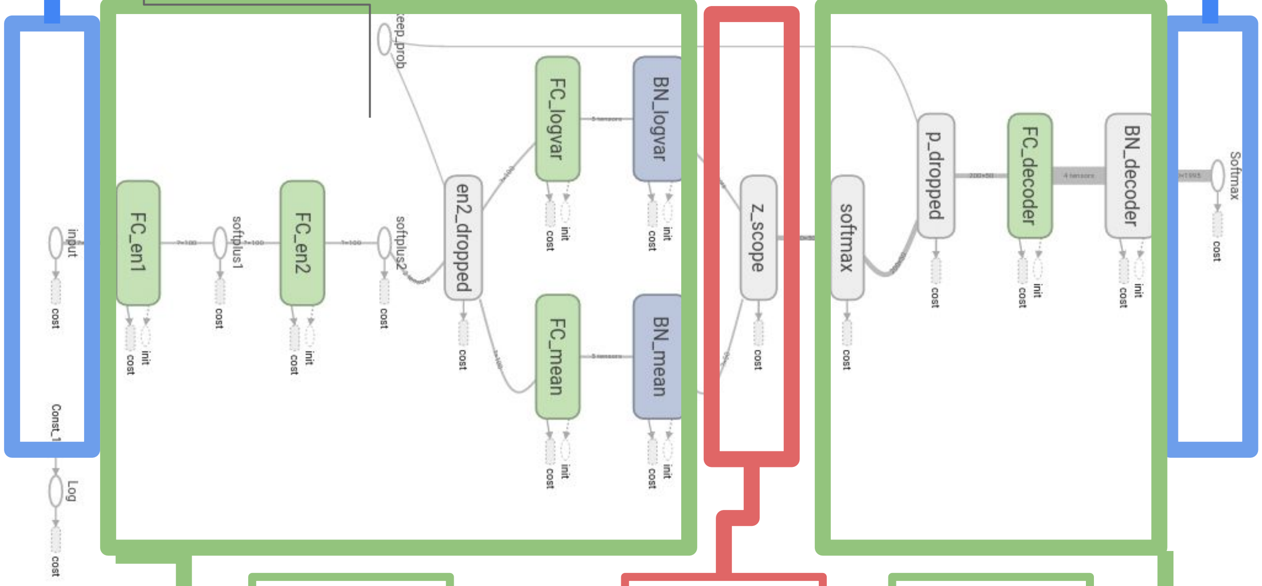
$$w_n | \beta_{z_n}, \theta \sim \text{Multinomial}(1, \sigma(\beta_{z_n} \theta))$$

Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Approach: ProdLDA

Input: document-word counts

Output: topic-word scores

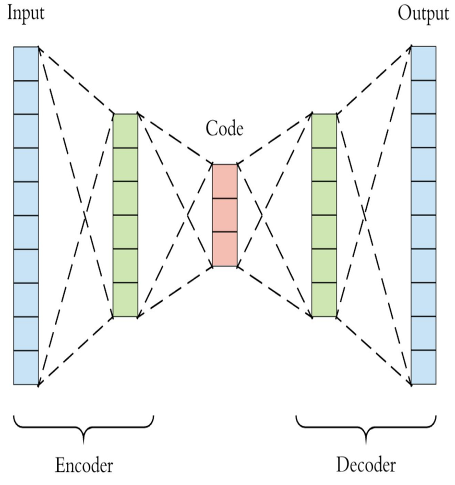


Encoder

“Topics”

Decoder

Traditional Autoencoder



Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Dataset

- **Dataset:** 20 Newsgroups data set (English articles)
- **Processing:** removed stop words, characters
- **Features:** Word count vector (features) for each document (independent observation).
    - Resulting in 18,745 documents and a vocabulary of 1995 words

Dataset: Lang, Ken. Newsweeder: Learning to filter netnews. Machine Learning Proceedings. Morgan Kaufmann, 1995. 331-339, 1995.

Evaluation metrics

- **Perplexity Score** - Measures the fit of the log-likelihood relative to a set of test articles. High log-likelihoods imply good fit.
  - Perplexity is inversely related log-likelihood ➔ lower values are better

$$\text{perplexity}(\text{test docs}) = \exp\left(\frac{-L(\text{test docs})}{\#\text{ of words}}\right)$$

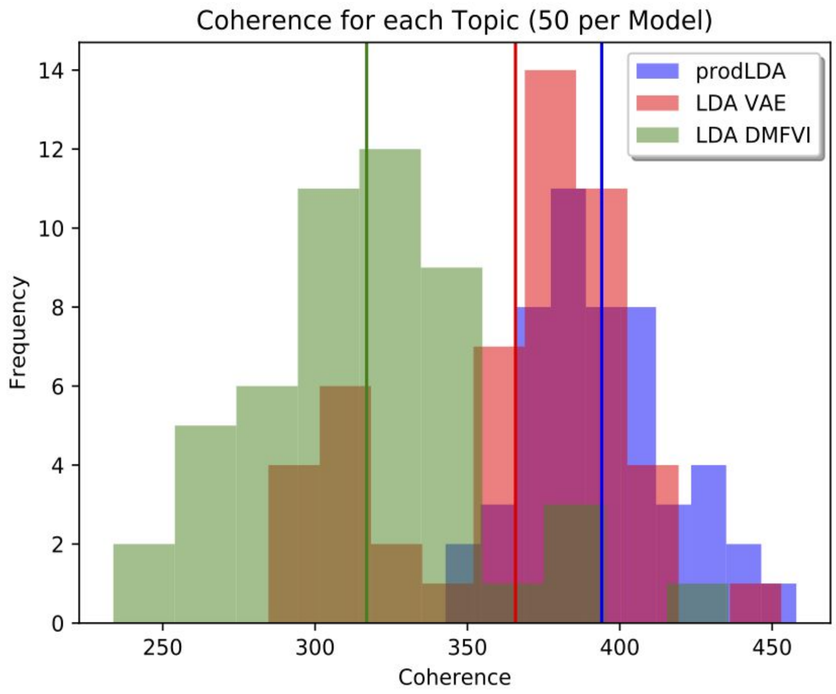
- **Topic Coherence** - Measure of how similar words are within a topic, or “intrinsic value”
  - Ideally, we want this value to be high

$$\text{coherence}(\text{test docs}) = \sum_{i<j} \log\left(\frac{p(w_i, w_j)}{p(w_i) \cdot p(w_j)}\right)$$

Results I: Quantitative Measures

Model	**Perplexity Score**
prodLDA	1153.7977
LDA VAE	1130.1957
LDA DMFVI (Traditional LDA topic model with variational inference solution)	777.7551

Measure of model fit
“Lower values are better”



Measure of intrinsic value
“Higher values are better”

Results II: ProdLDA Topic Keywords

Examples of ProdLDA model results with 50 Topics - Top 10 keywords

Topic (human inferred)	**Keywords**
Technology/encryption	*'anonymous', 'widget', 'ripem', 'visual', 'privacy', 'ftp', 'entry', 'int', 'binary', 'char'*
Sports/baseball	*'braves', 'hitter', 'pitcher', 'defensive', 'season', 'team', 'pitch', 'puck', 'deserve', 'fan'*
Religion/ Christianity	*'god', 'jesus', 'doctrine', 'satan', 'christ', 'revelation', 'worship', 'truth', 'christian', 'holy'*
Middle East	*'lebanese', 'israel', 'village', 'arab', 'arabs', 'lebanon', 'israeli', 'militia', 'turks', 'muslim'*

Conclusion

- Topic modeling via VAEs has more intrinsic value than traditional LDA.
- Relative performance of models consistent with results from Srivastava & Sutton (2017).

Challenges

- Results are sensitive to hyperparameters (E.g., Dirichlet hyperparameters)
- Some details for precise reproducibility are unavailable (I.e., hyperparameter values and selection process)

Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.