On the Interpretability of Discrete Set Representations: A Bayesian Perspective

Yao Fu
Department of Computer Science
Columbia University
yao.fu@columbia.edu

Abstract

We investigate the interpretability of the discrete set representation with a case study of the Latent Bag of Words Model. We show that the unsupervised learning of the bag of words representation in paraphrase generation does not necessarily lead to an interpretable representation. We demonstrate this is due to the permutation-invariance of the NLL objective with regard to the representation. To break the symmetry and enforce ordering, we use the empirical prior to regularize the posterior. Using an empirical Bayes approach, viewing the empirical prior as the aggregated posterior, we anneal the prior to the posterior by controlling the level of aggregation, thus smoothly interpolating the learning scheme from posterior regularization to supervised learning, which results in a progressively increased interpretability. Our work offers a bayesian perspective of the linguistic interpretability of a discrete set representation, and show how to learn an interpretable representation efficiently.

1 Introduction

Consider the following generative process for paraphrase generation: we have a input sentence x, we want to generate its paraphrase y, a different sentence conveying the same meaning of x. We assume a latent bag of words (BOW), i.e. a latent set, z encoding the all word neighbors of x. Since the distribution of a set has combinatorial complexity, we approximate it by a base categorical distribution over the vocabulary $p_\phi(\tilde{z}|x)$, and assume $z = [z_1, z_2, ..., z_k]$, as a set containing k different words, is sample from the base distribution $p_\phi(\tilde{z}|x)$ k times without replacement. Then y is generated by the decoder $p_\theta(y|z)$.

$$x \sim p(x) \tag{1}$$
$$z \sim p_\phi(\tilde{z}|x) \quad \text{sample } k \text{ times without replacement} \tag{2}$$
$$y \sim p_\theta(y|z) \tag{3}$$

We strictly restrict the representation to this set z, i.e. we do not disentangle any other continuous representations from x [1] [2] . To learn an interpretable z, we could use the BOW z^* from y as weak supervision, so that x will be encoded into a distribution representing the word neighbors close to y.

$$\mathcal{L}_z = \mathbb{E}_{p(z^*|y)}[-\log p_\phi(\tilde{z}|x)] \tag{4}$$
$$\mathcal{L}_y = \mathbb{E}_{p(y^*|x)}[-\log p_\theta(y|z)] \tag{5}$$

[1] By contrast, the LBOW model from Fu et al. [3] disentangles the encoding of x to a continuous representation and a discrete BOW representation.

[2] Of course this will come at the cost of the predictive performance, i.e. the BLEU will drop. But in this work we focus on learning interpretable representation.

Preprint. Work in progress.

Input	what is the most credible evidence of alien life forms or ufos
BOW sample	s old get within n't make ever
Output	is there any evidence of alien life form
Reference	extraterrestrial life : what is the most undeniable evidence of ufo ever seen
Input	what is your favorite video game of all time
BOW sample	get world discontinuing reduce ways girl make person
Output	what are your favorite games of all time
Reference	what are your favorite video games of all time and why

Figure 1: Samples from a latent z trained in an unsupervised way. The BOW z seems to be completely random with no interpretability.

Encoder	Input	what is the most credible evidence of alien life forms or ufos
	True BOW (implicit)	alien life ufo credible evidence ever
	Permuted BOW (observed)	s old get within n't make ever ---- Encoder converge to any permutation
Decoder	True BOW (implicit)	alien life ufo credible evidence ever ---- Decoder recover the permutation implicitly
	Output	is there any evidence of alien life form
	Reference	extraterrestrial life : what is the most undeniable evidence of ufo ever seen
Encoder	Input	what is your favorite video game of all time
	True BOW (implicit)	favorite video game time what
	Permuted BOW (observed)	get world discontinuing reduce ways girl make person ---- Encoder converge to any permutation
Decoder	True BOW (implicit)	favorite video game time what ---- Decoder recover the permutation implicitly
	Output	what are your favorite games of all time
	Reference	what are your favorite video games of all time and why

Figure 2: The encoder may converge to any permutation of the vocabulary, thus losing the interpretability. The decoder is able to learn that permutation implicitly, so the predictive performance remains.

Grounding z to with lexical semantics gives us an interpretable z, as is demonstrated in Fu et al. [3]. Here we ask a further question: if we drop the loss $\mathcal{L}_z$ in equation 4, can we still learn an interpretable z that automatically converge to the BOW of y?

It is very intuitive to think that z should converge to the words with the most predictve performance, which consequently should be the words in y. However, experiment results does not support this intuition. As is demonstrated in figure 1, the words seem to be totally random and does not correlate to y at all while the decoder could still generate correct sentences. We would like to ask why this could happen.

2 The Permutation-Invariance of the NLL Objective

The reason is closely related to Locatello et al. [7], as they demonstrate in the continuous gaussian case, the reconstruction likelihood is invariant to a differentiable bijective mapping f (e.g. a rotation) of the gaussian representation.

$$\mathcal{L}_y = \mathbb{E}_{p(y^*|x)}[-\log p_\theta(y|z)] = \mathbb{E}_{p(y^*|x)}[-\log p_\theta(y|f(z))] \tag{6}$$

In such cases, an interpretable disentangled representation will become entangled, thus non-interpretable. As an example, think about a two-dimensional gaussian representation for MNIST with the first dimension encoding the width of the strike, and the second dimension encoding the rotation of the digits. In such cases, controlling seperate dimensions will lead to corresponding changes of the generated digits. However, if we rotate the coordinate axis by 45 degree, the representation remain the same, but the interpretability vanishes as the dimensions are entangled now.

A discrete analogy to the above gaussian rotation would be a permutation of the vocabulary. i.e. f is a one-to-one mapping that maps any word w_i to another word $w_{i'} = f(w_i)$. Previous works have shown that neural networks are able to learn a permutation matrix [10, 6]. As is demonstrated in figure 2, even if we could have a interpretable bag of words z, any permutation of the vocabulary would lead the words in z to be mapped to other non-interpretable words. But the decoder is still able to first recover the true words by learning the permutation matrix implicitly, then generate the target sentence, under the same NLL objective.

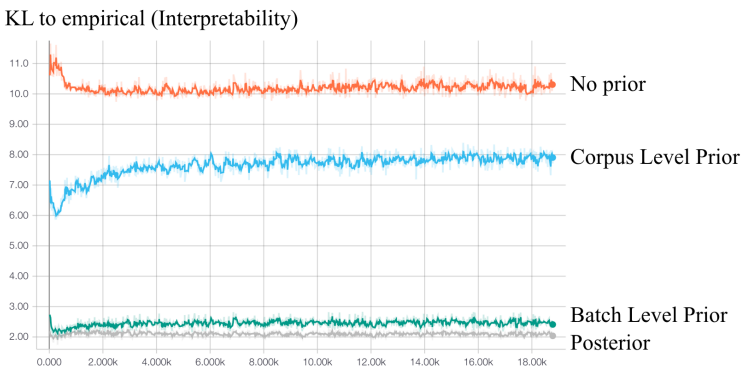

Figure 3: The KL divergence between the model posterior and the empirical posterior (training curve). As the loss anneals from prior (regularization) to posterior (supervised learning), the learned posterior becomes closer to the empirical posterior, the representation becomes more interpretable (see figure 5 for cases).

3 Related Works

The permutation-invariance property gives a discrete version of the conclusions of Locatello et al. [7]. It is one of the symmetry property that induces non-identifiability, discussed by many bayesian literature [9, 5, 2, 1, 7]. Generally, given a generative model $p(z|x) \propto p(x|z)p(z)$, the non-identifiability comes from: (1) the posterior $p(z|x)$ is symmetric for z (2) the prior $p(z)$ is symmetric for z. Since our recognition model is essentially a mixture model of words, it is inherently symmetric with respect to z, thus non-identifiable. To break symmetry and enforce ordering of the vocabulary, we could directly use supervision to the posterior, as is in Fu et al. [3]. In the most cases where we do not have labels, a typically bayesian approach it to use a ordered prior to regularize the posterior: $\mathcal{D}(p(z|x)||p(z))$ [8]. We investigate this approach in the following sections. We further show how to anneal the empirical prior to the labels of the posterior with a controlled batch size. We demonstrate that, as the regularization anneals from the empirical prior to the posterior labels, the model shows better ordering with a more restricted parameter space, thus improving interpretability progressively. Our approach gives a clear bayesian demonstration of interpretable representation learning: *the representation becomes more interpretable as the regularization anneals from prior to posterior*.

4 Enforce Ordering with the Empirical Prior

With no regualrization or supervision, the posterior $p(z|x)$ could converge to any permutation of z, i.e. assign high probability to non-interpretable words. A straigntforward approach to enforce ordering is to use labels for the posterior and do supervised learning:

$$\mathcal{L}_z = \mathbb{E}_{p(z^*|y)}[-\log p_\phi(\tilde{z}|x)] = \mathrm{KL}[p(z^*|x)||p_\phi(\tilde{z}|x)] - \mathcal{H}(p(z^*|x)) \tag{7}$$

As is shown in the above equation 7, optimizing the NLL loss is equivelent to optimizing the reverse KL divergence between the posterior and the empirical distribution. For many other machine learning tasks where we do not have label, but have knowledge about the prior distribution, the supervision loss could be replaced by a regualrization loss:

$$\mathcal{L}_{\text{prior}} = \mathrm{KL}[p_\phi(\tilde{z})||p(z)] \tag{8}$$

This is to say, we minimize the divergence between the aggregated posterior $p_\phi(\tilde{z})$ and the prior. Compared with supervised learning, here we use forward KL to encourage mode finding as the aggregated posterior, approximated at the batch level, should have less modes than the prior. Since the aggregated posterior is a categorical distribution over the vocabulary, we would like to use an empirical prior, i.e. the unigram distribution of the training set.

Table 1: Annealing from prior to posterior and the corresponding changes in learing scheme, objective, and the interpretability.

Learning scheme	Objective	Source of obejctive	Interpretability
Supervised Learning	$\mathrm{KL}(p_\theta(\tilde{z}\|x)\|\|p(z^*\|x))$	Posterior	Strong
Posterior Regularization	$\mathrm{KL}(p_\theta(\tilde{z})\|\|p(z^*))$	Batch level prior	Weak
Posterior Regularization	$\mathrm{KL}(p_\theta(\tilde{z})\|\|p(z^*))$	Corpus level prior	None
Unsupervised Learning	None	None	None

Figure 4: Controlling the level of aggregation with the batch size. As the batch size decreases, the interpretability increases.

5 Interpolate from Prior to Posterior: the Progressively Increased Interpretability

The empirical prior, although giving the word order at the corpus level, does not necessarily induce interpretability. Experiments shows similar inferred words (random-looking words) as the fully unsupervised case in figure 1. However, if we measure interpretability as the KL divergence between the model posterior and the empirical posterior (target BOW), we find out it is smaller compared with the unsupervised case, as is shown in figure 3. The reason we use the KL divergence as a metric of the interpretability can be understood by the following analysis: if the KL is 0, then the model perfectly predicts the target BOW, thus perfectly interpretable; if the KL is large, the model assign high probabilities to random-looking words; as the KL decreases, the model assigns higher probability to the target bow, the representation becomes more interpretable.

Due to the sparsity of this distribution and its long-tail nature, there are many words with the same frequency, thus the same probability. Words within the same probability are essentially non-identifiable from each other. So the empirical prior encourages a weak ordering by partitioning words into groups with different probabilities. Compared with prior, the bag of words of the target sentence y, induces an empirical posterior (the label) with strict ordering as it only promote the words in the target sentences.

A key observation is, the empirical prior is essentially the aggregated empirical posterior at the corpus level. If there is no aggregation at all, the learning scheme will reduce to supervised learning, and the model can learn the best interpretability. So we can progressively *controll the level of aggregation with the batch size*, as a way of interpolating between the prior and the posterior, thus controlling the interpretability of the learned representation. To achieve this, we construct a batch-level prior (instead of the corpus level prior), as the regualrization of the aggregated posterior. We should expect that as the batch sice decreases, the interpretability would increase. The annealing from the prior to the posterior, and the corresponding changes in learning scheme, objective, and interpretability, are listed in table 1.

Figure 3 shows that the interpretability increases as we change the learning scheme from unsupervised to supervised through regularized. Figure 4 shows that as we decrease the batch size, i.e. the level of aggregation, we anneal prior to posterior, and increase the interpretability. Figure 5 shows the cases of z as we adapt different learning schemes and how the interpretability evolves accordingly. Note that in the supervised case, although the word *band* is not interpretable with regard to the source sentence,

	Unsupervised / Regularized with Corpus level prior	Regularized with batch level prior	Supervised with posterior
Input	what is your favorite video game of all time		
BOW interpretable		favourite	favourite game videos all-time
BOW non-interpretable	get world discontinuing reduce ways girl	support rupees could sure general third	non band make
Output	what are your favorite games of all time	what is your favorite game and why	what is your favorite music video
Reference	what are your favorite video games of all time and why		

Figure 5: Cases of models trained with different level of regularization/ supervision. As the regularization anneal from prior to posterior, the interpretability of the representation increases, and becomes more aligned with human intuition.

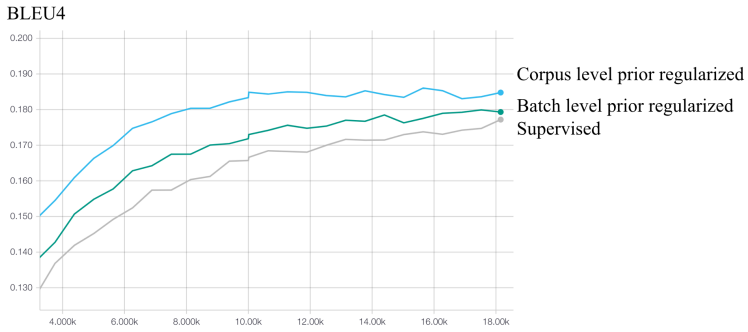

Figure 6: Tradeoff between interpretability and the performance. The stronger the representation is regualrized, the more interpretable it is, the less predictive performance it achieves.

it indeed influences the generated sentence with the word *music video*. From all the experiments, clearly we see that the representation becomes more interpretable as the regularization anneals from prior to posterior.

6 Tradeoff to Predictive Performance

An interpretable representation does not necessarily have good predictive performance, as is shown in figure 6. We see a clear tradeoff between the BLEU score and the interpretability. This result comforms to our general understanding of representation learning: a interpretable model requires more regularization and structure, thus losing the modeling flexibility, which results in decreased predictive performance [4, 11].

7 Discussion

Choice of the Prior. The reason that the empirical prior cannot strictly enforce the interpretable ordering is that there are many words with the same prior probability, thus non-identifiable. Alternatively, to break the symmetry at this level, we could use a prior with a finer structure, like a prior from a topic model. We leave this to future work.

Interpretability and Sparsity. In our scenario, only words that are neighbors of the source and target are interpretable. Without supervision, we should actually treat the infer index as a set of latent tags, i.e. discard its connection to words. But in this case it is hard to determine what these index represent because they are too spase. The interpretability is closely related to the sparsity, as a dense counterpart is the unsupervised NER, where the number latent tags are typically less than 10. In the dense case we could actually gather words with different tags and determine what they represent (names, locations, .etc). The reason the empirical prior cannot enforce an interpretable BOW is also related to sparsity, as there are many words have the same proability. We could also restrict the size of the latent vocabulary to decrease the sparsity from the model architecture, we skip this for now. Our conclusion is, under a BOW model architecture, to learn an interpretable discrete representation, one have to use the target BOW information as weak supervision. This conclusion should generalize to other discrete latent representation with sparsity as symmetry is a common problem in such models.

8 Conclusion

We analysis the unsupervised learning of interpretable discrete representaion with the LBOW model. We use an empirical bayes approach to smoothly interpolating the learning scheme from posterior regularization to supervised learning, resulting in a progressively increasing interpretability. Our conclusion aligns with the result of Locatello et al. [7], which tells that without inductive bias on model and date, the unsupervised learning of disentangled representaion is impossible. We further show that, in our BOW setting, due to the sparsity, to learn an interpretable representaion, one need explicit supervision for the latent representation.

References

[1] Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*, 2017.

[2] Michael Betancourt. Identifying bayesian mixture models. 2017. URL `https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html`.

[3] Yao Fu, Yansong Feng, and John P Cunningham. Paraphrase generation with latent bag of words. pages 13623–13634, 2019. URL `http://papers.nips.cc/paper/9516-paraphrase-generation-with-latent-bag-of-words.pdf`.

[4] Yoon Kim, Alexander M Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. Unsupervised recurrent neural network grammars. *arXiv preprint arXiv:1904.03746*, 2019.

[5] Erich L Lehmann and George Casella. Theory of point estimation. *Springer Science & Business Media*, pages 24–24, 2006.

[6] Scott W Linderman, Gonzalo E Mena, Hal Cooper, Liam Paninski, and John P Cunningham. Reparameterizing the birkhoff polytope for variational permutation inference. *arXiv preprint arXiv:1710.09508*, 2017.

[7] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[8] Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. *arXiv preprint arXiv:1812.02833*, 2018.

[9] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

[10] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.

[11] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Learning neural templates for text generation. *arXiv preprint arXiv:1808.10122*, 2018.