On the Effect of Isotropy on VAE Representations of Text

Anonymous ACL submission

Abstract

Injecting desired geometric properties into text representations has attracted a lot of attention. A property that has been argued for, due to its better utilisation of representation space, is isotropy. In parallel, VAEs have been successful in areas of NLP, but are known for their sub-optimal utilisation of the representation space. To address an aspect of this, we investigate the impact of injecting isotropy during training of VAEs. We achieve this by using an isotropic Gaussian posterior (IGP) instead of the ellipsoidal Gaussian posterior. We illustrate that IGP effectively encourages isotropy in the representations, inducing a more discriminative latent space. Compared to vanilla VAE, this translates into a much better classification performance, robustness to input perturbation, and generative behavior. Additionally, we offer insights about the representational properties encouraged by IGP.[1]

1 Introduction

In recent years, with the success facilitated by pre-trained representations across various NLP tasks, more attention has been placed on studying and utilising the geometric properties of learned representations. A phenomena that has been studied more recently in this direction is anisotropy (Ethayarajh, 2019), indicating a suboptimal property where the learned embeddings only utilise a small subset of the representation space. Various methods have been proposed to rectify this and encourage the representations to be more discriminative and or to exploit the representation dimensions more effectively (Liu et al., 2021; Gao et al., 2021; Li et al., 2020a; Su et al., 2021; Mu and Viswanath, 2018).

In parallel, Variational Autoencoders (VAEs) (Kingma and Welling, 2014) have been widely used in various areas of NLP, from representation learning for downstream tasks (Li et al., 2020b; Wei and Deng, 2017), to generation (Prokhorov et al., 2019; Bowman et al., 2016), and semi-supervised learning (Zhu et al., 2021; Choi et al., 2019; Yin et al., 2018; Xu et al., 2017). In recent years, most of the developments around VAEs have focused on avoiding the posterior collapse (Bowman et al., 2016) which leads to learning sub-optimal representations (Havrylov and Titov, 2020; Fu et al., 2019; Li et al., 2019; Dieng et al., 2019; He et al., 2019; Higgins et al., 2017; Yang et al., 2017; Bowman et al., 2016). Despite the success of these techniques, a non-collapsed VAE still utilises the representation space sub-optimally (Prokhorov et al., 2019; He et al., 2019; Burda et al., 2016), as very commonly the learned representations do not fully utilise the latent space to encode information.

In this paper we bridge between the two lines of research by injection isotropy in the latent space of VAEs. Such property could be encouraged by using an Isotropic Gaussian Posterior (IGP) which involves a simple modification of VAEs. An Isotropic Gaussian distribution, $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, is similar to vanila VAE's posterior with the exception that all dimensions share the same unified variance. Tying the variances would encourage encoder of VAEs towards the extremes where *all* dimensions are either active or inactive.[2]

Our experimental findings indicate that, compared to vanilla VAE: The use of IGP is effective in both increasing dimension activation and injecting isotropy in the learned representation space. We observe that isotorpy results in a more discriminative representation space which is much more suited for classification tasks and robust to input perturbation. Our generative experiment for sentence completion suggests that the VAE trained with IGP is more capable of maintaining semantic cohesiveness.

[1] Our code, data, and scripts will be released at publication.

[2] A dimension u is defined to be active if $A_u = \mathrm{Cov}_{\mathbf{x}}(\mathbb{E}_{u\sim q(u|\mathbf{x})}[u])$ is larger than 0.01, where Cov denotes covariance (Burda et al., 2016).

2 Isotropic Gaussian Posterior (IGP)

Variational Autoencoder (VAE). Let $\mathbf{x}$ denote datapoints in data space and $\mathbf{z}$ denote latent variables in the latent space, and assume the datapoints are generated by the combination of two random processes: The first random process is to sample a point $\mathbf{z}^{(i)}$ from the latent space in VAEs with prior distribution of $\mathbf{z}$, denoted by $p(\mathbf{z})$. The second random process is to generate a point $\mathbf{x}^{(i)}$ from the data space, denoted by $p(\mathbf{x}|\mathbf{z}^{(i)})$. VAE uses a combination of a probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and decoder $p_\theta(\mathbf{x}|\mathbf{z})$, parameterised by ϕ and θ, to learn this statistical relationship between $\mathbf{x}$ and $\mathbf{z}$. VAE is trained by maximizing the lower bound of the logarithmic data distribution $\log p(\mathbf{x})$, called evidence lower bound (ELBO), $\mathcal{L}(\phi, \theta; \mathbf{x})$:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{x}|\mathbf{z}))] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

The first term of objective function is the expectation of the logarithm of data likelihood under the posterior distribution of z. The second term is KL-divergence, measuring the distance between the recognition distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior distribution $p(\mathbf{z})$ and can be seen as a regularisation.

In the presence of auto-regressive and powerful decoders, a common optimisation challenge of training VAEs in text modelling is called posterior collapse, where the learned posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$, collapses to the prior $p(\mathbf{z})$. Several strategies have been proposed to alleviate this problem (Bowman et al., 2016; Havrylov and Titov, 2020; Fu et al., 2019; He et al., 2019). In this work, we follow Prokhorov et al. (2019), $\mathcal{L}(\phi, \theta; \mathbf{x})$:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{x}|\mathbf{z}))] - |\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|$$

where C is a positive real value which represents the target KL-divergence term value. We set $\beta = 1$ to make sure the weights of the two terms balance, noting that it acts as a Lagrange Multiplier (Boyd and Vandenberghe, 2004). This also has an information-theoretic interpretation, where the placed the KL term is seen as the amount of information transmitted from a sender (encoder) to a receiver (decoder) via the message ($\mathbf{z}$) (Alemi et al., 2018) and the usage of C can control this channel capacity. This can help us to make a fair comparison between DGP and IGP when VAEs are under the same encoder capacity constraint.

VAE with Isotropic Gaussian Posterior. A common behaviour of VAEs is the presence of inactive representation units across the entire dataset, causing the number of utilised dimensions to be even far smaller than the number of potential generative factors behind any real-world dataset. The soft ellipsoidal representation space of VAEs is known to lead to less representative mean vectors (Bosc and Vincent, 2020). We illustrate that encouraging isotropy (i.e., tying the variance of dimensions on the posterior) will avoid the aforementioned issue since the encoder of VAEs would be forced to either use all dimensions or none and the learned latent space is soft spherical. In the Gaussian case, this corresponds to using an Isotropic Gaussian, a subclass of diagonal Gaussian distribution $\{\mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})|\boldsymbol{\mu} \in \mathbf{R}^n, \sigma \in \mathbf{R}^+\}$, as the posterior. Tying the variances in IGP imposes a different pathological pattern by encouraging AU to reach the maximum (i.e., representation dimension).

Additionally, the use of IGP allows the estimation of variance more accurately. Suppose we have N samples with the same posterior. For a K-dimension diagonal Gaussian posterior, we will have an estimate of variance with standard deviation approximately $\hat{\sigma}_k^2\sqrt{\frac{2}{N}}$ for each dimension k, whereas for an isotropic Gaussian posterior, we will have a unified estimation of variance with standard deviation approximately $\hat{\sigma}^2\sqrt{\frac{2}{NK}}$, where $\hat{\sigma}_k^2$ and $\hat{\sigma}^2$ denote the estimates of the variance. Moreover, with K different $\hat{\sigma}_k^2$ estimates, a few may differ substantially from their best values by chance.[3]

3 Experiments

We trained our models on Yahoo Question and DBpedia (Zhang et al., 2015) which have (100K/10K/10K, 12K, 10) and (140K/14K/14K, 12K, 14) for (sentences in training/dev/test, vocabulary size, classes), respectively. We use the VAE architecture of (Bowman et al., 2016) and concatenate the latent code with word embedding at every timestamp as the input of the decoder. For VAE with IGP, we just produce one variance value and assign it to be the variance of posterior for all dimensions. At decoding phase, we use greedy decoding. The dimensions for word embedding, encoder-decoder LSTMs, and latent code are (200, 512, 32). Three different values of C are used on each dataset to explore the impact of the amount of information transmitted by the code. We also

[3] It is worth noting that IGP is not a solution for posterior collapse, and our experimental findings are not specific to the chosen technique for avoiding the collapse (i.e., our preliminary experiments with KL-annealing exhibit similar findings reported in this paper).

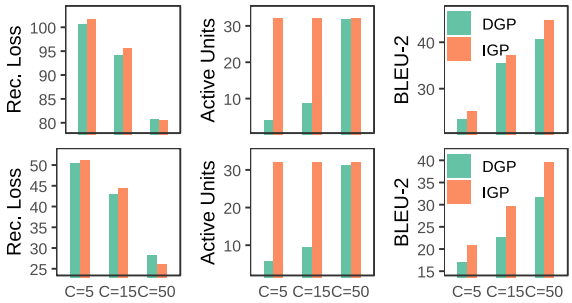

Figure 1: Results are calculated on the test set (average of 3 runs reported) of (top) DBpedia Corpus and (bottom) Yahoo Question (Zhang et al., 2015). AU is bounded by the dimensionality of z (32).

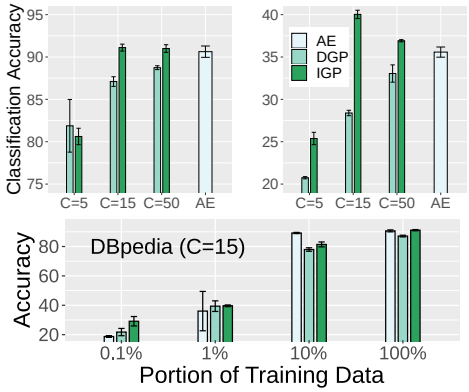

Figure 2: Classification accuracy on DBpedia (top-left) and Yahoo (top-left) with and without the isotropic Gaussian posterior (IGP) under different C values. Also, classification accuracy for $C = 15$ trained on various portion of DBpedia (bottom). Results are reported as mean and std across 3 VAE encoders.

adopt Autoencoder (AE) as a baseline.[4] All models are trained from 3 random starts for 20 epochs and 128 batch size using Adam (Kingma and Ba, 2015) with learning rate 0.0005.

We compare the choice of isotropic Gaussian posterior (IGP) with vanilla diagonal Gaussian posterior (DGP) on various grounds, from reconstruction loss and unit activation (§3.1) to downstream classification task, sample efficiency, robustness, and generation (§3.2), and distributional properties of the induced representations (§3.3).

3.1 Basic Results

Figure 1 reports the reconstruction loss, active units (AU; Burda et al., 2016) and BLEU-2 (Papineni et al., 2002) for $C = 5, 15, 50$. KL in all cases match the set target C. We observe the C constraint can effectively control the KL-divergence to the set level. The reconstruction loss generally drops with the increase of C. We observe the same pattern for DGP and IGP. Additionally, while DGP struggles, IGP can activate all dimensions (e.g., AU for C=5 on DBpedia are 4 and 32 for DGP and IGP, respectively). This translates into IGP reaching a significantly higher BLEU. For more results, including autoencoder, see *Appendix*.

3.2 Classification and Generation

Classification. We trained a classifier on top of the frozen encoders of DGP and IGP and use the mean vector representations as a feature to train the classifier. For the classifier, we used a 2-hidden-layer MLP with 128 neurons and ReLU activation function at each layer. We trained 10 randomly initialised classifiers and used the mean of classification accuracy as the final accuracy. Figure 2 (top) reports the results. Overall, the representations of most VAEs with IGP lead to a significant improvement of classification accuracy compared to vanilla VAEs. In the only exception (i.e, $C = 5$ on DBpedia), two models have comparable results with no model having any statistically significant advantage. We attribute this to having a more representative mean which is encouraged by IGP. One notable thing is that DGP does not perform as good as AEs regardless of C choice, whereas IGP ($C = 15, 50$) achieve similar and better classification accuracy on DBpedia and Yahoo Question, respectively.

We adopted few-shot setting to compare sample efficiency of both VAEs, by using 0.1%, 1% and 10% of training data of DBpedia to them with $C = 15$ and do classification on the test set as before. Accuracy scores are reported in Figure 2 (bottom) with IGP exhibiting a better sample efficiency. For instance, the mean accuracy gap at 0.1% is quite significant being above 7 points, and VAE gets the gap down to 4 points at 100% (still significant).

We further investigated the robustness of the learned representations to perturbation via applying word dropout on sentences by randomly deleting 30% of words in a sentence, and repeating the classification experiment. IGP with accuracies of (83.5, 34) outperforms both DGP (76.4, 24.1) and AE (83.1, 30.7) on (DBpedia, Yahoo). We speculate this to be an indication of information overlap across dimensions of the representations at higher AU, offering a better recovery of information in the presence significant perturbation.

[4] We also tried Importance Weighted Autoencoder (IWAE;Burda et al. (2016)) as another baseline commonly used in image domain. This model yields KL-collapse which is non-trivial to address given its objective function.

ORIGINAL	the carnegie library in unk washington is a building from 1911 . it was listed on the national register of historic places in 1982 .
IMPUTED	the carnegie library in unk washington · · ·
DGP	the carnegie library in unk washington is a unk (unk ft) high school in the unk district of unk in the province of unk in the unk province of armenia .
IGP	the carnegie library in unk washington was built in 1909 . it was listed on the national register of historic places in unk was designed by architect john unk .

Table 1: Word imputation experiment.

	DBpedia		Yahoo	
	Sample	Mean	Sample	Mean
DGP	0.72	0.62	0.72	0.63
IGP	0.76	0.77	0.78	0.76
AE	0.087		0.059	

Table 2: Isotropy score of mean and samples for DBpedia and Yahoo test sets (trained with $C = 15$).

Generation. We imputed %75 of words of a sentence from the test set of DBpedia, fed it to VAE encoder and reconstructed the sentence from its latent code using IGP and DGP in Table 1. IGP successfully recovers the type of the mentioned object and complete the imputed sentence with a similar structure, whereas DGP fails to do so (another example is provided in *Appendix*).

3.3 Properties of Representations.

Isotropy Score. We quantitatively approximate the isotropy score (Mu and Viswanath, 2018),

$$IS(\mathcal{V}) = \frac{\min_{m \in \mathcal{M}} \sum_{v \in \mathcal{V}} \exp(m^\intercal v)}{\max_{m \in \mathcal{M}} \sum_{v \in \mathcal{V}} \exp(m^\intercal v)},$$

where $\mathcal{V}$ is the matrix of representations (i.e., of samples or mean vectors of posteriors), and $\mathcal{M}$ is the set of eigen vectors of $\mathcal{V}^\intercal\mathcal{V}$. As observed in Table 2, compared to DGP, IGP has a significantly larger IS on both means and samples. Interestingly, given that dimensions are independently modeled via univariate Gaussians, both VAEs outperform the Autoencoder counterparts.

Visualization. We visualize the learned representation space of DGP and IGP for DBpedia, using t-sne (van der Maaten and Hinton, 2008), in Figure 3 (Bottom). As illustrated in the right plot, the clusters of classes in IGP has less overlap among classes compared with DGP (left). Additionally, we use the Mapper[5] algorithm (Singh et al., 2007) to visualise the highest density region (HDR) (Hyndman, 1996) of the mean vectors for DGP and IGP. HDR cuts the overall density space to form latent spaces that contain above a threshold probability mass (i.e., ≥ 0.05 with minimum samples ≥ 2 per latent space). The output of the mapper is a graph, where each component in the graph corresponds to a set of nearby points forming a high density space. The connectivity of the graph reflects some topological properties of the sampling space (darker colors indicate higher density). As observed in Figure 3 (Top), the HDR of DGP posterior means forms a single component whereas IGP forms 9 disconnected components indicating a more discriminative characteristics of its mean vectors, echoing earlier results in better accuracy in the classification setting (§3.2).

[5] github.com/scikit-tda/kepler-mapper

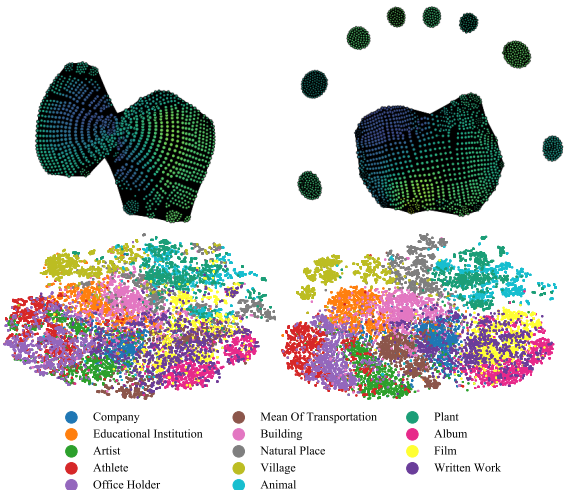

Figure 3: Visualisations of the mean representations of posterior on DBpedia test set for $C = 15$. **Left:** DGP; **Right:** IGP. **Top:** HDR; **Bottom:** T-SNE.

4 Conclusion

We proposed Isotropic Gaussian Posteriors (IGP) as the means for encouraging isotropy in the latent space induced by VAEs. The injection of isotropy addressed a sub-optimal behaviour of VAEs by activating more dimensions of the representation and encouraging a more discriminative latent space. Our experiments illustrated a significant improvement of classification performance and robustness to input perturbations with IGP. We also observed, in the sentence completion task, that VAE trained with IGP is more capable at maintaining semantic cohesiveness. Our ongoing work suggests the representation utilisation achieved by IGP has the potential to be exploited towards representational properties such as disentanglement.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. 2018. Fixing a broken ELBO. In *International Conference on Machine Learning*, pages 159–168. PMLR.

Tom Bosc and Pascal Vincent. 2020. Do sequence-to-sequence VAEs learn global features of sentences? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4296–4318, Online. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, pages 10–21, Berlin, Germany. ACL.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge University Press.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. 2016. Importance weighted autoencoders. In *4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings*, San Juan, Puerto Rico.

Jihun Choi, Taeuk Kim, and Sang-goo Lee. 2019. A cross-sentence latent variable model for semi-supervised text sequence matching. In *ACL*, pages 4747–4761.

Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. 2019. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 2397–2405, Naha, Okinawa, Japan. PMLR.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Serhii Havrylov and Ivan Titov. 2020. Preventing posterior collapse with levenshtein variational autoencoder. *CoRR*, abs/2004.14758.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Lagging inference networks and posterior collapse in variational autoencoders. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, Toulon, France.

Rob J Hyndman. 1996. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, CA, USA.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*, Banff, AB, Canada.

Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. 2019. A surprisingly effective fix for deep latent variable modeling of text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3603–3614, Hong Kong, China. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020b. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Victor Prokhorov, Ehsan Shareghi, Yingzhen Li, Mohammad Taher Pilehvar, and Nigel Collier. 2019. On the importance of the Kullback-Leibler divergence term in variational autoencoders for text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019*, pages 118–127, Hong Kong. Association for Computational Linguistics.

Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. 2007. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *CoRR*, abs/2103.15316.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Liangchen Wei and Zhi-Hong Deng. 2017. A variational autoencoding approach for inducing cross-lingual word embeddings. In *IJCAI*, pages 4165–4171.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *AAAI*, pages 3358–3364.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3881–3890, Sydney, NSW, Australia. PMLR.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In *ACL*, pages 754–765.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Yi Zhu, Ehsan Shareghi, Yingzhen Li, Roi Reichart, and Anna Korhonen. 2021. Combining deep generative models and multi-lingual pretraining for semi-supervised document classification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 894–908, Online. Association for Computational Linguistics.

A Full Results

We report detailed reconstruction loss, KL-divergence, active units and results of BLEU and ROUGE scores on reconstructed test set in Table 3.

B Generation Results

Another example of generation results is provided in Table 4.

C Full Robustness Results

We report full results for various C settings in Table 5.

D Posterior Shape

To understand the impact of isotropy on the aggregated posterior, $q_\phi(z) = \sum_{x \sim q(x)} q_\phi(z|x)$, we obtain unbiased samples of z by sampling an x from data and then $z \sim q_\phi(z|x)$, and measure the log determinant of covariance of the samples ($\log\det(\mathrm{Cov}[q_\phi(z)])$) as well as the mean of the samples to measure $||\mu||_2^2$. Table 6 reports these for $C = 15$. We observe that $\log\det(\mathrm{Cov}[q_\phi(z)])$ is significantly lower for IGP indicating a sharper approximate posteriors.

		Rec.	KL	AU	BLEU-2/4	ROUGE-2/4
DBpedia	AE	$66.32_{\pm 0.11}$	-	$32.0_{\pm 0.0}$	$40.96_{\pm 0.25}/27.57_{\pm 0.17}$	$35.87_{\pm 0.19}/23.78_{\pm 0.07}$
	$C = 5$, DGP	$100.65_{\pm 0.08}$	$5.09_{\pm 0.01}$	$4.0_{\pm 0.0}$	$23.47_{\pm 0.79}/14.00_{\pm 0.47}$	$20.85_{\pm 0.69}/10.19_{\pm 0.34}$
	$C = 5$, DGP	$101.73_{\pm 0.31}$	$5.04_{\pm 0.01}$	$32.0_{\pm 0.0}$	$25.09_{\pm 0.55}/14.83_{\pm 0.22}$	$22.29_{\pm 0.17}/10.47_{\pm 0.10}$
	$C = 15$, DGP	$94.16_{\pm 0.19}$	$15.06_{\pm 0.04}$	$8.7_{\pm 0.9}$	$35.35_{\pm 0.49}/22.37_{\pm 0.31}$	$30.54_{\pm 0.43}/17.41_{\pm 0.16}$
	$C = 15$, IGP	$95.52_{\pm 0.08}$	$15.08_{\pm 0.05}$	$32.0_{\pm 0.0}$	$37.23_{\pm 0.27}/24.47_{\pm 0.11}$	$34.19_{\pm 0.12}/19.32_{\pm 0.06}$
	$C = 50$, DGP	$80.65_{\pm 0.53}$	$50.02_{\pm 0.04}$	$31.7_{\pm 0.5}$	$40.54_{\pm 0.21}/26.95_{\pm 0.19}$	$35.19_{\pm 0.25}/22.13_{\pm 0.24}$
	$C = 50$, IGP	$80.58_{\pm 0.04}$	$50.15_{\pm 0.04}$	$32.0_{\pm 0.0}$	$44.79_{\pm 0.30}/30.72_{\pm 0.17}$	$39.91_{\pm 0.12}/25.40_{\pm 0.08}$
Yahoo Question	AE	$17.64_{\pm 0.28}$	-	$32.0_{\pm 0.0}$	$42.88_{\pm 0.51}/32.86_{\pm 0.58}$	$41.63_{\pm 0.58}/31.67_{\pm 0.67}$
	$C = 5$, DGP	$50.58_{\pm 0.06}$	$5.14_{\pm 0.01}$	$5.7_{\pm 0.5}$	$17.07_{\pm 0.71}/6.04_{\pm 0.25}$	$10.96_{\pm 0.41}/1.50_{\pm 0.06}$
	$C = 5$, IGP	$51.24_{\pm 0.01}$	$5.06_{\pm 0.03}$	$32.0_{\pm 0.0}$	$20.91_{\pm 0.03}/8.07_{\pm 0.03}$	$14.48_{\pm 0.13}/2.21_{\pm 0.02}$
	$C = 15$, DGP	$43.00_{\pm 0.12}$	$15.06_{\pm 0.04}$	$9.3_{\pm 1.2}$	$22.62_{\pm 0.37}/10.81_{\pm 0.21}$	$16.04_{\pm 0.32}/4.76_{\pm 0.08}$
	$C = 15$, IGP	$44.43_{\pm 0.05}$	$15.20_{\pm 0.12}$	$32.0_{\pm 0.0}$	$29.76_{\pm 0.06}/14.99_{\pm 0.08}$	$23.11_{\pm 0.17}/6.94_{\pm 0.12}$
	$C = 50$, DGP	$28.29_{\pm 0.40}$	$50.00_{\pm 0.19}$	$31.3_{\pm 0.9}$	$31.78_{\pm 0.73}/20.47_{\pm 0.70}$	$27.14_{\pm 0.85}/15.07_{\pm 0.77}$
	$C = 50$, IGP	$26.18_{\pm 0.19}$	$50.15_{\pm 0.08}$	$32.0_{\pm 0.0}$	$39.68_{\pm 0.20}/27.49_{\pm 0.31}$	$35.73_{\pm 0.40}/22.57_{\pm 0.55}$

Table 3: Results are calculated on the test set. We report mean value and standard deviation across 3 runs. Rec, AU, and PPL denote reconstruction loss, number of Active Units and estimated perplexity, respectively. DGP, and IGP denote diagonal Gaussian posteriors and isotropic Gaussian posteriors, respectively. C is the target KL value.

ORIGINAL	st. marys catholic high school is a private roman catholic high school in phoenix arizona . it is located in the roman catholic diocese of phoenix .
IMPUTED	st. marys catholic high school is $\cdots$
DGP	st. marys catholic high school is a unk - unk school in unk unk county new jersey united states . the school is part of the unk independent school district .
IGP	st. marys catholic high school is a private roman catholic high school in unk california . it is located in the roman catholic diocese of unk .

Table 4: Word imputation experiment on DBpedia test set.

	DBpedia	**Yahoo**
AE	$83.09_{\pm 0.81}$	$30.68_{\pm 0.32}$
$C = 5$, DGP	$68.77_{\pm 4.76}$	$18.47_{\pm 0.50}$
$C = 5$, IGP	$72.46_{\pm 1.00}$	$22.28_{\pm 0.30}$
$C = 15$, DGP	$76.42_{\pm 1.18}$	$24.08_{\pm 0.28}$
$C = 15$, IGP	$\mathbf{83.49_{\pm 0.63}}$	$\mathbf{33.99_{\pm 0.20}}$
$C = 50$, DGP	$79.18_{\pm 0.38}$	$28.29_{\pm 1.02}$
$C = 50$, IGP	$83.04_{\pm 0.70}$	$32.16_{\pm 0.19}$

Table 5: Results on robustness. Classification accuracies after 30% Proportion of reserved label information on test set and robustness.

	CBT	DBpedia	Yahoo
DGP	$[0.11, -0.41]$	$[0.11, -0.63]$	$[0.10, -0.51]$
IGP	$[0.11, -1.48]$	$[0.12, -6.22]$	$[0.08, -4.86]$

Table 6: Reports $[\ ||\mu||_2^2\ ,\ \log\det(\mathrm{Cov}[q_\phi(z)])\ ]$.