Polarized-VAE: Proximity Based Disentangled Representation Learning for Text Generation

Vikash Balasubramanian[1]*, Ivan Kobyzev[2], Hareesh Bahuleyan[2], Ilya Shapiro[3], Olga Vechtomova[1]
[1]University of Waterloo, [2]Borealis AI, [2]University of Windsor
{v9balasu,ovechtom}@uwaterloo.ca
ivan.kobyzev@borealisai.com
hareeshbahuleyan@gmail.com
ishapiro@uwindsor.ca

Abstract

Learning disentangled representations of real world data is a challenging open problem. Most previous methods have focused on either fully supervised approaches which use attribute labels or unsupervised approaches that manipulate the factorization in the latent space of models such as the variational autoencoder (VAE), by training with task-specific losses. In this work we propose polarized-VAE, a novel approach that disentangles selected attributes in the latent space based on proximity measures reflecting the similarity between data points with respect to these attributes. We apply our method to disentangle the semantics and syntax of a sentence and carry out transfer experiments. Polarized-VAE significantly outperforms the VAE baseline and is competitive with the state-of-the-art approaches, while being more a general framework that is applicable to other attribute disentanglement tasks.

1 Introduction

Learning representations of real word data using deep neural networks has accelerated research within a number of fields including computer vision and natural language processing (Zhang et al., 2018). Previous work has advocated for the importance of learning *disentangled representations* (Bengio et al., 2013; Tschannen et al., 2018).

Although attempts have been made to formally define disentangled representations (Higgins et al., 2018), there is no widely accepted definition of disentanglement. However, the general consensus is that a disentangled representation should separate the distinct factors of variations that explain the data (Bengio et al., 2013). Intuitively, a greater level of interpretability can be achieved when different independent latent units are used to encode different independent ground-truth attributes of the data (Burgess et al., 2018).

However, recovering and separating all the distinct factors of variation in the data is an extremely challenging problem. For complex real world datasets, there may not be a way to separate each factor of variation into a single dimension in the learned fixed size vector representation. An easier problem would be to separate complex factors of interest into distinct subspaces of the learned representations. For instance, the representation of text could be separated into *content* and *style* which could allow for style transfer.

For disentangling factors of variation, a commonly used approach is based on adversarial training (John et al., 2019). However, adversarial methods pose optimization challenges and may lead to unstable training. An alternative strategy used by Locatello et al. (2018) builds on the objective of decreasing mutual information or total correlations. A limitation of such approaches is that estimation of mutual information for continuous variables is not straightforward, especially when dealing with high dimensional spaces (Hjelm et al., 2019).

In this work, we explore an orthogonal approach and propose the polarized-VAE to disentangle the latent space into subspaces corresponding to different factors of variation. We control the relative location of representations in a particular latent subspace, based on the similarity of their respective data points according to the corresponding criterion. This encourages similar points to be grouped together and dissimilar points to be farther away from each other in that subspace. Figuratively, we polarize the latent subspaces, hence the name.

In summary, the main contributions of this paper are three-fold: (1) We propose a general framework for learning disentangled representations. Even though we test our method on an NLP task, the underlying concept is very general and can be applied to other domains such as computer vision; (2) We provide a method for disentanglement that does

not rely on adversarial training or specialized multitask losses; (3) We demonstrate an application of our method by disentangling the latent space into subspaces corresponding to syntax and semantics. Such a setting can be used to perform controlled text decoding such as generating a paraphrase with a desired sentence structure.

2 Related Work

Unsupervised disentanglement of underlying factors using the Variational Autoencoder (VAE, Kingma and Welling (2013)) framework has been studied by Higgins et al. (2017); Kim and Mnih (2018). However, Locatello et al. (2018) show that completely unsupervised disentanglement of the underlying factors may be impossible without supervision or inductive biases. Unsupervised disentanglement for text has been shown to be especially difficult, but attempts have been made to leverage it for controllable text generation (Xu et al., 2019).

Most previous work on supervised disentanglement for text has focused on adversarial training. (John et al., 2019; Yang et al., 2018). Recently, the task of disentangling the semantics and syntax of text into distinct subspaces has received attention from researchers. Chen et al. (2019b) use several multitask losses such as paraphrase loss and word position loss in sentence VAE models to encourage learning of separate semantic and syntactic information in the latent space. Bao et al. (2019) use adversarial training and make use of syntax trees along with specific multitask losses to disentangle semantics and syntax.

We propose the polarized-VAE approach where disentanglement is achieved through distance based learning. In contrast to previous approaches, our method does not require the use of several multitask losses or adversarial training, both of which can result in optimization challenges. At the same time, we don't need precise attribute labels, but simply proxy labels based on the concept of similarity.

3 Background

VAEs serve as a foundation for many natural language tasks including natural language generation and representation learning. It uses a probabilistic encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to encode a sentence $\boldsymbol{x}$ into a latent variable $\boldsymbol{z}$, and a probabilistic decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ that attempts to reconstruct the original sentence $\boldsymbol{x}$ from its latent representation $\boldsymbol{z}$. The objective is to minimize the following loss function:

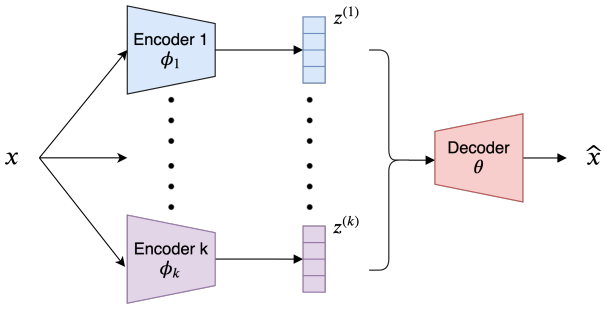

Figure 1: polarized-VAE model architecture

$$\mathcal{L}_{\text{vae}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} \tag{1}$$

where $\mathcal{L}_{\text{rec}} = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$ is the sentence reconstruction loss and $\mathcal{L}_{\text{kl}} = \mathcal{D}_{\text{kl}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$ is the Kullback-Leibler (KL) divergence loss. The KL term ensures that the approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is close to the prior $p(\boldsymbol{z})$, which is typically assumed to be the standard normal $\mathbf{N}(0, \mathbb{I})$; λ_{kl} is a hyperparameter that controls the extent of KL regularization.

4 Approach

The idea behind our approach is to impose additional proximity regularization on the latent subspaces learned by VAEs.

4.1 Disentanglement into Subspaces

We assume that we have a collection of criteria $C = \{c_1, ..., c_k\}$, based on which we wish to disentangle the latent space $\boldsymbol{z}$ of the VAE into k subspaces: $\boldsymbol{z} = [\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(k)}]$. Here $\boldsymbol{z}^{(i)}$ denotes the latent subspace corresponding to the criterion c_i.

In this paper, we focus on the case where the latent space is disentangled into semantics (c_1) and syntax (c_2), i.e., $k = 2$.

4.2 Supervision based on Similarity

We assume that we have information (possibly noisy) about pairwise similarities of the input sentences. Given a pair of sentences, the similarity information can be either a binary label (if both sentences belong to the same class) or an integer or continuous scalar variable (e.g., edit distance). In this work, the similarity criterion is a binary label:

$$\text{Sim}(\boldsymbol{x}_i, \boldsymbol{x}_j|c) = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are similar} \\ & \text{w.r.t. the criterion } c \in C \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

In our case, the two criteria for disentanglement are semantics (c_1) and syntax (c_2). We use this

additional information to regularize the encoders of the VAE, by adding the proximity based loss functions on the latent subspaces, $D(\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_j^{(1)}|c_1)$ and $D(\boldsymbol{z}_i^{(2)}, \boldsymbol{z}_j^{(2)}|c_2)$.

4.3 Training Method and Proximity Function

Extending the traditional VAE approach, we have a set of RNN-based encoders parameterized by ϕ_c that encode the approximate posteriors $q_{\phi_c}(\boldsymbol{z}^{(c)}|\boldsymbol{x})$. Given two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we denote the proximity of their encodings in the latent subspace by $D(q_{\phi_c}(\boldsymbol{z}^{(c)}|\boldsymbol{x}_i), q_{\phi_c}(\boldsymbol{z}^{(c)}|\boldsymbol{x}_j))$.

We experiment with multiple forms of proximity functions (see Section 5.5) and found the cosine distance between the samples to perform the best, i.e.,

$$D(q_{\phi_c}(\boldsymbol{z}|\boldsymbol{x}_i), q_{\phi_c}(\boldsymbol{z}|\boldsymbol{x}_j)) = d_c(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{1}{2}\left(1 - \frac{\boldsymbol{z}_i \boldsymbol{z}_j}{||\boldsymbol{z}_i|| ||\boldsymbol{z}_j||}\right) \quad (3)$$

Based on the above distance, we add a regularization term to the VAE loss function as follows. For each example $(\boldsymbol{x}, c)$, we have a positive sample $\boldsymbol{x}_p$ and m negative samples $\boldsymbol{x}_{n_1}, ..., \boldsymbol{x}_{n_m}$, such that $\text{Sim}(\boldsymbol{x}, \boldsymbol{x}_p|c) = 1$ and $\text{Sim}(\boldsymbol{x}, \boldsymbol{x}_{n_j}|c) = 0$; $j \in \{1, ..., m\}$:

$$\mathcal{L}_c = \max(0, 1 + d_c(\boldsymbol{z}, \boldsymbol{z}_p) - \frac{1}{m}\sum_{j=1}^{m} d_c(\boldsymbol{z}, \boldsymbol{z}_{n_j})) \quad (4)$$

This regularization function can be viewed as a max-margin loss over the proximity function. The final objective then becomes

$$\mathcal{L} = \mathcal{L}_{\text{vae}} + \sum_{c=1}^{C} \lambda_c \mathcal{L}_c \quad (5)$$

The overall model architecture of polarized-VAE is illustrated in Figure 1.

5 Experiments

To demonstrate the effectiveness of polarized-VAE in obtaining disentangled representations, we carry out semantics-syntax separation of textual data, using the Stanford Natural Language Inference (SNLI, Bowman et al. (2015)) dataset.

5.1 Reconstruction and Sample Quality

We evaluate our model on reconstruction and sample quality to ensure that the distance regularization used does not adversely impact the reconstruction or the sampling capabilities of the standard VAE.

For this purpose, we compare our model and the standard VAE on two metrics: reconstruction BLEU (Papineni et al., 2002) and the Forward Perplexity (PPL)[1] (Zhao et al., 2018) of the generated sentences obtained by sampling from the model's latent space. As seen in Figure 2 there is a clear trade-off as expected between reconstruction quality and sample quality. Overall, polarized-VAE performs slightly better than standard VAE and this indicates that the proximity-based regularization does not inhibit the model capabilities.

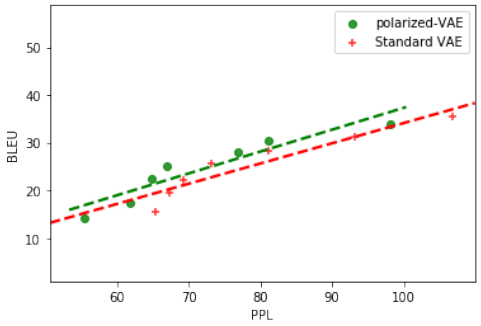

Figure 2: Comparing trade-off between BLEU and PPL for Standard VAE and polarized-VAE, with different KL coefficients λ_{kl}.

5.2 Controlled Generation and Transfer

We follow the work of Chen et al. (2019a); Bao et al. (2019) and analyze the performance of controlled generation by evaluating syntax transfer in generated text. Given two sentences, $\boldsymbol{x}_{\text{sem}}$ and $\boldsymbol{x}_{\text{syn}}$ we wish to generate a sentence that combines the semantics of $\boldsymbol{x}_{\text{sem}}$ and the syntax of $\boldsymbol{x}_{\text{syn}}$ using the following procedure:

$$\boldsymbol{z}_{\text{sem}} \sim q_{\phi_1}(\boldsymbol{z}^{(1)}|\boldsymbol{x}_{\text{sem}}) \;\; ; \;\; \boldsymbol{z}_{\text{syn}} \sim q_{\phi_2}(\boldsymbol{z}^{(2)}|\boldsymbol{x}_{\text{syn}})$$

$$\boldsymbol{z} = [\boldsymbol{z}_{\text{sem}}, \boldsymbol{z}_{\text{syn}}] \;\; ; \;\; \boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$$

Following the evaluation methodology of Bao et al. (2019), we measure transfer based on (1) semantic content preservation for the semantic subspace and (2) the tree edit distance (Zhang and Shasha, 1989) for the syntactic subspace.

We consider a subset of pairs of sentences from the SNLI dataset (1000 pairs) for evaluation. We want the generated sentence to be close to $\boldsymbol{x}_{\text{sem}}$ and different from $\boldsymbol{x}_{\text{syn}}$ in terms of semantics, which is measured using BLEU scores. We also report the difference to indicate the strength of transfer denoted by ΔBLEU.

[1] PPL is computed using the KenLM toolkit (Heafield et al., 2013)

Model	BLEU			TED			
	$\boldsymbol{x}_{\text{sem}}{}^{\uparrow}$	$\boldsymbol{x}_{\text{syn}}{}^{\downarrow}$	ΔBLEU$^{\uparrow}$	$\boldsymbol{x}_{\text{sem}}{}^{\uparrow}$	$\boldsymbol{x}_{\text{syn}}{}^{\downarrow}$	ΔTED$^{\uparrow}$	ΔGM$^{\uparrow}$
Standard VAE	4.75	4.67	0.08	13.70	13.60	0.10	0.28
Bao et al. (2019)	**13.74**	6.15	7.59	**16.19**	13.10	**3.08**	4.83
polarized-VAE	10.78	0.92	**9.86**	14.09	11.67	2.42	**4.88**
polarized-VAE (`wo`)	9.82	0.84	8.98	14.12	11.65	2.47	4.71
polarized-VAE (`len`)	10.10	**0.76**	9.34	12.68	**11.44**	1.44	3.67
polarized-VAE (`wo`, `len`)	9.41	0.87	8.54	12.65	11.48	1.17	3.16

Table 1: Results of syntax transfer generation on SNLI dataset. Bao et al. (2019) report TED after multiplying by 10, we report their score after correcting for it.

Additionally, we would like the generated sentence to be syntactically similar to $\boldsymbol{x}_{\text{syn}}$ and different from $\boldsymbol{x}_{\text{sem}}$, which is measured by per sentence average Tree Edit Distance (TED). We also report ΔTED to indicate the strength of the syntax transfer. Finally, we use the Geometric Mean of ΔBLEU and ΔTED to report a combined score ΔGM . We also provide qualitative examples of our transfer experiments in the Appendix.

Our default variant of polarized-VAE uses the entailment labels from SNLI dataset as a proxy for semantic similarity. For syntactic similarity we threshold the differences in tree edit distance (of syntax parses) as a proxy for syntactic similarity. We also evaluate two other variants of our model on this task. In the model variant polarized-VAE (`wo`) (see Table 1) we use BLEU scores as a heuristic proxy for estimating semantic similarity, while keeping the syntactic training unchanged. We also experiment with heuristics for syntax in polarized-VAE (`len`) where we use length as a heuristic proxy for syntax, while still making use of the ground truth similarity labels for the semantic training. Finally we combine these two heuristics in polarized-VAE (`wo`, `len`) which can be viewed as an unsupervised variant that does not make use of any labels or syntax trees.

Our model outperforms the VAE baseline on all metrics (Table 1). In comparison to (Bao et al., 2019), our model is much better at ignoring the semantic information present in $\boldsymbol{x}_{\text{syn}}$ during syntax transfer, as evidenced by our lower BLEU scores w.r.t. $\boldsymbol{x}_{\text{syn}}$. On the other hand, we perform slightly worse in BLEU w.r.t. $\boldsymbol{x}_{\text{sem}}$. Our model does a better job at matching the syntax of the sentence $\boldsymbol{x}_{\text{syn}}$ as indicated by the lower TED score w.r.t. $\boldsymbol{x}_{\text{syn}}$.

5.3 Disentanglement

There is a possibility that the two latent spaces may encode similar information. But that is likely to happen only if the attributes themselves are highly correlated (e.g., if we want to disentangle syntax from length). For such cases, even existing methods based on adversarial disentanglement (John et al., 2019) may fail to completely separate out correlated information.

However, if the attributes are different enough (or ideally independent) for e.g., syntax and semantics, this is less problematic. Note that we apply our proximity loss independently to each of the subspaces (i.e., leaving the other space(s) untouched for a given input). This encourages the semantic encoder to encode semantically similar sentences close together and dissimilar ones far apart in the semantic space (same applies for the syntax encoder).

We empirically compute correlations between the semantic and syntax latent vectors for 1000 test sentences, to check whether the two encoders learn similar information.

By feeding 1000 sentences from the test set to the Polarized-VAE, we obtain their corresponding semantic (z_{sem}) and syntax (z_{syn}) latent vectors. We then empirically compute the correlation between z_{sem} and z_{syn}. To analyze the level of similarity of information represented in z_{sem} and z_{syn}, we report the maximum absolute correlation (max across all pairs of dimensions) and also the mean absolute correlation. A higher value of correlation would indicate that there is more overlapping information learnt by the semantic and syntactic encoders. As illustrated in Table 2, the analysis indicates that the semantic and syntax latent vectors in Polarized-VAE encodes less correlated information than Baseline VAE (due to the proximity-based regularization). This demonstrates that the 2 latent spaces learned by our model encode sufficiently different information.

Model	Max Abs Corr↓	Mean Abs Corr↓
Baseline-Vae	0.62	0.1
Polarized-Vae	**0.25**	**0.05**

Table 2: Maximum Absolute Correlation and Mean Absolute Correlation between the semantic and syntactic latent vectors.

5.4 Human Evaluation

In addition to the above experiments, we carried out a human evaluation study for comparing the generated outputs. The test setup is as follows - we provide as input two sentences, x_{sem} and x_{syn} to the model; we wish to generate a sentence that combines the semantics of x_{sem} and the syntax of x_{syn}. We asked 5 human annotators to evaluate the outputs from the 3 models: Baseline-VAE, Polarized-VAE and the model from (Bao et al., 2019).

Each annotator was shown the input sentences (x_{sem} and x_{syn}) and the outputs from the 3 models (randomized so that the evaluator is unaware of which output corresponds to which model). They were then asked to pick the one best output for each of the following three criteria: 1) Semantic transfer (level of semantic similarity with respect to x_{sem}), 2) Syntactic transfer (level of syntactic similarity with respect to x_{syn}) and 3) Fluency.

We obtained annotations on 100 test set examples. To aggregate the annotations, we used majority voting with manual tie breaking to find the best model for each test example (and for each test criteria). In Table 3, we report the percentage of instances for which each of the models were chosen as the best model, according to human evaluations under the 3 criteria: semantics transfer, syntax transfer and fluency.

Model	Semantics	Syntax	Fluency
Baseline-Vae	11	11	**43**
(Bao et al., 2019)	24	**58**	31
Polarized-Vae	**65**	31	38

Table 3: Human Evaluation scores on Semantics Syntax and Fluency reported as percentages.

We note that polarized-VAE is better at semantic transfer and worse at syntactic transfer in comparison to (Bao et al., 2019). The human evaluation results are consistent with the results from automatic evaluation metrics, where polarized-VAE scores higher on Δ BLEU (indicator of semantic transfer strength) and (Bao et al., 2019) is better at Δ TED (indicator of syntax transfer strength). With respect to fluency criterion, polarized-VAE ranks higher than (Bao et al., 2019). However, the most fluent sentences are produced by the baseline VAE. We hypothesise this to be due to the presence of additional regularization terms in the loss functions of both (Bao et al., 2019) and polarized-VAE, which in turn affects the fluency of their generated text (due to the deviation from the reconstruction objective).

5.5 Proximity Functions

We considered several proximity functions over the posterior distributions: KL-Divergence, Hellinger Distance, Maximum Mean Discrepancy (MMD), and the generalized Jensen-Shannon divergence that has a closed form solution for Gaussian Distributions (Nielsen, 2019). We also considered using the cosine distance over just the means of the Gaussian posteriors.

The best results however were obtained with the cosine distance between the samples as our proximity function. It is symmetric, bounded, and continuous and also has an intuitive geometrical interpretation. We noticed that the unbounded divergence functions caused instability issues during training and could easily lead to the loss function diverging to large values as a result of the negative sampling procedure involved.

6 Conclusion and Future Work

In this paper, we proposed a general approach for disentangling latent representations into subspaces using proximity functions. Given a pair of data points, a predefined similarity criterion in the original input space determines how close or how far they are positioned in the corresponding latent subspace, which is modelled via a proximity function.

We apply our approach to the task of disentangling semantics and syntax in text. Our model, polarized-VAE, significantly outperforms the VAE baseline and is competitive with the state-of-the-art approach while being more general as we do not use specific multitask losses or architectures to encourage preferring semantic or syntactic information. Our methodology is orthogonal to the multitask learning approaches by Chen et al. (2019b) and Bao et al. (2019) and hence, can be naturally combined with their methods.

For future work, we would like to investigate this approach on disentanglement applications out-

side of NLP. Another interesting research direction would be to further explore suitable proximity functions and identify their properties that could facilitate disentanglement.

References

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. Association for Computational Linguistics (ACL).

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019a. Controllable paraphrase generation with a syntactic exemplar. In *Proc. of ACL*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019b. A multi-task approach for disentangling syntax and semantics in sentence representations. 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 2453–2464.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. 2018. Towards a definition of disentangled representations.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmssan, Stockholm Sweden. PMLR.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Frank Nielsen. 2019. On a generalization of the jensen-shannon divergence and the js-symmetrization of distances relying on abstract means. *arXiv preprint arXiv:1904.04017*.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Tschannen, Olivier Bachem, and Mario Lucic. 2018. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*.

Peng Xu, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. On variational learning of controllable representations for text without supervision.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7287–7298. Curran Associates, Inc.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18:1245–1262.

Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li. 2018. A survey on deep learning for big data. *Information Fusion*, 42:146–157.

Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *35th International Conference on Machine Learning, ICML 2018*, pages 9405–9420. International Machine Learning Society (IMLS).

A Model and Training Details

Both the semantic and syntactic encoders are bidirectional LSTMs (Hochreiter and Schmidhuber, 1997) with hidden state size of 128. followed by two single hidden layer feedforward networks to parametrize the Gaussian loc (μ) and scale (σ) parameters similar to standard VAE formulations used by (Bao et al., 2019). The latent space dimensions were taken to be $dim(z^1) = 64$ and $dim(z^2) = 16$. The decoder is a unidirectional LSTM with a hidden size of 128.

We adopt the standard tricks for VAE training including dropout and KL annealing followed by (Bowman et al., 2016). We anneal both semantic and syntactic KL weights (λ_{kl}) upto 0.3 (5000 steps) using the same `sigmoid` schedule.

We train the model for 30 epochs in total using the ADAM optimizer (Kingma and Ba, 2014) with the default parameters and a learning rate of 0.001.

B Proximity Functions

We provide the results for the proximity functions that we have used in our experiments

Metric	ΔBLEU$^{\uparrow}$	ΔTED$^{\uparrow}$	ΔGM$^{\uparrow}$
Cosine Distance	**9.86**	**2.42**	**4.88**
Hellinger Distance	4.12	0.86	1.42
MMD	5.21	1.17	1.91
KL Divergence	4.32	0.75	1.28
JS Divergence	5.81	1.46	2.33

Table 4: Comparison of different proximity functions we used in our experiments.

We note that since there is no closed form expression for the JS divergence between two Normal Random variables we used the generalized JS Divergence proposed by (Nielsen, 2019).

C Transfer Examples

We provide qualitative examples of our transfer experiments, where we generate a sentence with the semantics of x_{sem} and the syntactic structure of x_{syn} in Table 5. We also provide the sentences generated by a standard-VAE for comparison.

$\boldsymbol{x}_{\mathbf{sem}}$	$\boldsymbol{x}_{\mathbf{syn}}$	**polarized-VAE**	**standard-VAE**
A man works near a vehicle.	A woman showing her face from something to her friend.	A man directing traffic on a bicycle to an emergency vehicle.	A woman works on a loom while sitting outside.
A family in a party preparing food and enjoying a meal.	Man reading a book.	A person enjoying food.	A man plays his guitar.
Two young boys are standing around a camera outdoors.	Three kids are on stage with a vacuum cleaner.	Two young boys are standing around a camera outdoors.	Two people are standing on a snowy hill.
There are a group of people sitting down.	They are outside.	There are people.	They are outside
a woman wearing a hat and hat is chopping coconuts with machete.	The person is in a blue shirt playing with a ball.	a woman with a hat is hanging upside down over utensils.	A girl in a pink shirt and elbow pads is swirling bubbles.
The young girl and a grownup are standing around a table , in front of a fence.	A guy stands with cane outdoors.	The young girl is outside.	The little boy is doing a show.
A person is sleeping on bed.	A man and his son are walking to the beach , looking for something.	A man and a child sit on the ground covered in bed with rocks.	A man is wearing blue jeans and a blue shirt walking.
The men and women are enjoying a waterfall.	A dog is holding an object.	The man and woman are outdoors.	The two men are working on the roof.
a man dressed in uniform.	There is a man with a horse on it.	A man dressed in black clothing works in a house.	A man dressed in black and white holding a baby.

Table 5: Examples of transferred sentences that use the semantics of x_{sem} and syntax of x_{syn}