Effective data augmentation for sentence classification using one VAE per class

Piedboeuf, Frédéric and **Langlais, Philippe**
DIRO, RALI
Université de Montréal
Montréal, Québec, Canada

Abstract

In recent years, data augmentation has become an important field of machine learning. While images can use simple techniques such as cropping or rotating, textual data augmentation needs more complex manipulations to ensure that the generated examples are useful. Variational auto-encoders (VAE) and its conditional variant the Conditional-VAE (CVAE) are often used to generate new textual data, both relying on a good enough training of the generator so that it doesn't create examples of the wrong class. In this paper, we explore a simpler way to use VAE for data augmentation: the training of one VAE per class. We show on several dataset sizes, as well as on four different binary classification tasks, that it outperforms other generative data augmentation techniques.

1 Introduction

Data augmentation (DA) has been shown to be an efficient technique for deep neural networks (Ramirez Rochac et al., 2019), consisting in artificially creating new labelled examples and therefore inflating the size of the dataset. While a great number of data augmentation techniques have been developed in recent years, most of these require external data, either in a database form (Wei and Zou, 2019), pretrained embeddings (Marivate and Sefara, 2020), or neural networks (Wu et al., 2019). Even if these techniques help improve classifiers, they cannot be used in a variety of contexts, most notably in artificial intelligence for rare languages, where there simply isn't enough training data to pre-train efficient models (Feldman and Coto-Solano, 2020), or where the collection of labelled data for fine-tuning is difficult. The ethical importance of considering smaller communities and other languages in machine learning research has been noted repeatedly (Bender et al., 2021; Fazelpour and De-Arteaga, 2022), and consequently, it is important to develop techniques that can work for languages other than English.

In this paper, we take a special look at generative models for data augmentation, and more specifically at Variational Auto-Encoders (VAE), for binary classification tasks. Generative algorithms such as the VAE are especially interesting because they do not require external data and, therefore, can be used on a variety of domains. In particular, we show that using one separate VAE per class and generating new data by random sampling of the latent space is a very efficient way of inflating the size of a dataset. To the best of our knowledge, this technique has been considered only once before, in the context of balancing unbalanced datasets, and with somewhat disappointing results (Qiu et al., 2020). Contrary to them, we study the technique in the context of pure data augmentation, on more datasets and various starting sizes of datasets, and conclude that it is an efficient technique. We test two variations of this method, with or without sharing parameters between the VAEs, and show that both variations are equally efficient.

We compare this technique to two others generative techniques that are commonly used in the DA literature, mainly using VAEs as a paraphrasing system, or using a Conditional VAE (CVAE) for directly generating from a given class. We show that they perform consistently worse than the separate VAE approach. We surmise that this is because, while we lose the ability to perform style transfer between classes, we gain a stronger guarantee that the generated examples will not be of an erroneous class. We furthermore compare to another technique that showed excellent results, CBERT (Wu et al., 2019), and show using one VAE per class for DA outperforms it as well.

This paper is organized as follows: Sections 2, 3, and 4 present respectively a brief summary of VAEs, the relevant literature to our paper, and the description of our various generative DA methods. Then, Section 5 and 6 present our datasets and the results of our experiments. Finally, we present a

discussion of the results in Section 7, the broader implications of our work in Section 8, and conclude in Section 9.

2 Variational Auto Encoders

VAEs are generative models introducing a latent variable $z \sim p(\mathbf{z})$, where data points (in our case, sentences) are assumed to be generated following $p(\mathbf{x}|\mathbf{z})$. VAEs follow the general auto-encoder structure, with an encoder that predicts the parameters associated with the latent distribution, most often a diagonal Gaussian, and a decoder that takes samples from the latent distribution and transforms them into sentences. At training time, sampling is done using the reparametrization technique (Kingma and Welling, 2014), which samples from $\mathbf{z} = \mu + \sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. This allows the gradient to flow through μ and σ. The training objective is the ELBO, or Evidence Lower Bound:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \quad (1)$$

where ϕ represents the parameters of the encoder, and θ, the parameters of the decoder. The CVAE is a modification of the VAE where we instead assume that $\mathbf{x}$ is generated conditioned on both the class of the example and the latent space, formally $p(\mathbf{x}|\mathbf{z}, \mathbf{c})$, where $\mathbf{c}$ is the class. In practice, this means that we give the class to the model while both training and generating, so that we can directly generate examples from the desired class (Yan et al., 2016).

In order to use the VAE for text, the standard solution is to replace both the encoder and the decoder with Recurrent Neural Networks (RNN). We show a simple example of a textual VAE model in Figure 1.

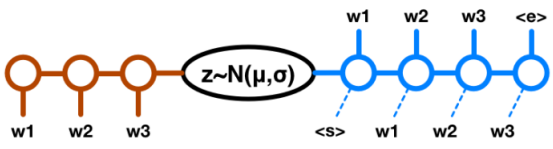


Figure 1: A textual VAE model. Image from Semeniuta et al. (2017).

This, however, often comes with the problem of KL-collapse, where the VAE relies entirely on the powerful decoder to generate sentences and collapses the latent distribution to a single point. To prevent this, we resort to two common techniques, namely KL-annealing, which slowly brings the strength of the KL-term in the ELBO from 0 to 1, and word dropout, which randomly masks words for the decoder while training (Bowman et al., 2016). Both of these techniques encourage the VAE to rely on the latent distribution.

3 Related Work

Data augmentation has been studied extensively (Feng et al., 2021; Shorten and Khoshgoftaar, 2019). We focus our review on studies targeting data augmentation for sentence classification, as well as works using the VAE for DA, which are the most pertinent studies for our work.

The general principle of data augmentation is fairly direct: starting with a dataset of fixed size, we generate new (hopefully) diverse data that we can then use to feed the classifier, diminishing generalization error. While DA for images can perform simple operations, such as cropping or rotating the images (Wang et al., 2017), things are a little more complicated for textual data, since operations on words (such as replacing or deleting words), have a chance of generating examples of the wrong class. As such, a lot of the focus is put on class-coherence, which simply means that generated examples have the class we want them to.

Many approaches have been developed for data augmentation for sentence classification, the simplest ones consisting in performing word-level operations. For example, the authors of (Marivate and Sefara, 2020) use W2V (Mikolov et al., 2013) for replacing random words in a sentence, picking words that are close to the original one in a pretrained embedding space. In (Xiang et al., 2021), replacements are made using the part of speech information to ensure they only replace content words. More operations can also be considered, as in EDA (Wei and Zou, 2019; Liesting et al., 2021), which performs four operations on the sentences: replacement by a synonym, deletion of words, swapping words, and inserting synonyms of words at random positions. In EDA, synonyms are found by using WordNet (Miller, 1995) instead of pretrained embeddings.

Substitutions can also be made using pretrained neural networks with a masked word task, such as an LSTM (Kobayashi, 2018), or conditional-BERT (CBERT) (Wu et al., 2019), which is a simple extension of BERT where we prepend the class of

the example to the sentence, in the hope that it will learn to generate class-coherent outputs.

Generative models, and most notably VAEs, are frequently used in data augmentation, often using a conditional model to directly generate from a given class (Zhuang et al., 2019; Malandrakis et al., 2019; Wang et al., 2020; Rizos et al., 2019). This, however, relies on the assumption that the CVAE is learning to correctly generate from the wanted class. While CVAEs have made tremendous progress in recent years, it remains a difficult task for textual CVAE.

The use of unconditional VAEs is also commonly explored. In Islam et al. (2021), the boundary decision is found in the latent space so that generation can be made on either side of it, depending on the class we want to generate data for. A popular technique is to use the VAE as a paraphrase machine (Mesbah et al., 2019; Malandrakis et al., 2019), encoding and decoding the data points and relying on the stochasticity of the system to generate variations of the examples. This technique, however, is also based on the hope that the algorithm will learn to generate from the same class as the input example. Finally, VAEs have also been used to generate new data that is not class dependent, such as spoken utterances for spoken language understanding (Yoo et al., 2019).

Qiu et al. (2020) attempt, similarly to our paper, to use one VAE per class for balancing data. They compare it to CVAE as well as EDA. They show, however, that generative algorithms barely outperform sampling strategies when balancing the data, on two datasets. Contrary to them, we study the use of one VAE per class in a pure data augmentation context, with already balanced classes, and show that it does work better than other algorithms. We also test all the algorithms on several dataset sizes and on four datasets, providing analysis and examples of why and how it works. Last but not least, we show that it works with modern classification algorithms (in this case, BERT), and show that it works better than state-of-the-art data augmentation.

4 VAE for data augmentation

In this paper, we revisit the technique of Qiu et al. (2020), which trains one VAE per class for generating new examples for unbalanced data, but use it for general data augmentation. We test two variants of this, one with parameter sharing (sharing of the encoder and embeddings), denoted VAE-Linked, and one where everything is separated, denoted VAE-Sep. Concretely, in VAE-Sep we initialize and train m VAEs, where m is the number of classes, and each VAE is trained only on the data of one class, that is to say on all examples $x_i|y_i \epsilon c_k$, where y_i is the class of the example x_i, and c_k is the class number k. For VAE-Linked, we initialize one VAE model with m decoders, each decoding only examples of one class. When generating, we first choose the class we want to generate from, select the corresponding VAE, and generate by randomly sampling from its latent space. This process is illustrated in Figure 2.

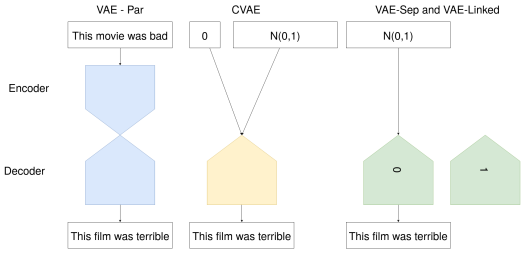


Figure 2: Illustration of the generation process of a negative movie review with the three VAE models tested in this paper. In VAE-Par, the new examples are created by passing sentences through the VAE and relying on its stochasticity to create paraphrases. In CVAE, the sentence is generated by sampling from the latent space and conditioning the decoding on the negative class. Finally, in VAE-Sep, we train two different VAEs, one for the negative sentences and one for the positive ones. We generate negative examples by sampling from the latent space and using the negative decoder. In VAE-Linked, the generation happens identically to VAE-Sep, but while training, we share the encoder and the embeddings between classes. Positive examples are generated in a similar way but passing a positive sentence, the positive label, or using the positive decoder.

We compare this to two popular DA techniques using VAEs: conditional generation (CVAE) and paraphrasing (VAE-Par). While both VAE-Sep and VAE-Linked lose the advantage of style transfer between classes and also have fewer data points to learn to generate sentences, it also greatly reduces the risk of generating examples of the wrong class.

As mentioned, generative models for textual DA are interesting due to the fact that they do not need external data to work. As such, we do not use pretrained embeddings in the RNNs of the VAEs[1]. We

[1]While outside the scope of this study, preliminary experiments with GLoVE embeddings (Pennington et al., 2014), in place of embeddings initialized randomly, did not bring a

discuss in Section 9 ideas for future research in a context where the use of external data is encouraged.

5 Datasets

We compare the models on four binary sentence classification tasks. We use movie critics classification with SST-2 (Socher et al., 2013), detection of fake news articles with the FakeNews dataset[2], detection of ironic tweets with the dataset Irony (Van Hee et al., 2018), and detection of subjective questions with Subj (Bjerva et al., 2020). Characteristics of the datasets are summarized in Table 1.

Dataset	# Ex.	Av. Sent. Length
SST-2	6920	19.3
FakeNews	12799	12.5
Irony	2683	14.3
Subj	13990	5.6

Table 1: Characteristics of the datasets used in this study. The average length represents the average number of words, when tokenized at white spaces.

6 Experiments

Because we are interested in domains where data is limited, we sample portions of the datasets to act as our initial training sets. We test 4 different datasets sizes: 500, 1000, 1500, and 2000. When sampling, we make sure to balance classes. We run each experiment 15 times and report the average results. As our baseline, we train BERT without data augmentation.[3]

We report in Table 2 the basic hyperparameters for all four variational algorithms. We use a GRU with 1 layer as our RNN for both the encoder and decoder, and a sigmoid function $\sigma(x)$ for annealing the KL-divergence from 0 to 1, which takes in the parameters x_0, controlling where $\sigma(x) = 0.5$, and a parameter k which controls the strength of the slope. While we fix x_0 at 15 (out of 30 epochs), we calculate it in practice according to the total number of batches, taking a step of annealing at every batch. Doing so allows us to keep a fixed k parameter, which makes the final KL-weight lower

$x0$	15
k	0.0025
Batch Size	32
Latent Size	15
Hidden Size	2048
Nb Epoch	30
Dropout	0.5
Word dropout	0.6
Nb layers GRU	1

Table 2: Basic hyperparameters used for all three variational algorithms

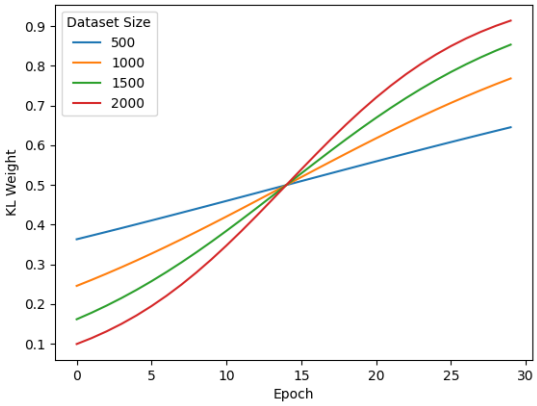


Figure 3: Annealing of the KL weight for various dataset sizes.

for a smaller dataset size (which are more prone to KL-collapse) and achieves 1 for a larger dataset size, as shown in Figure 3. We found that this setting worked well for all four sizes tested. To simulate a real world environment, we fine-tune to generate good sentences before testing on the test set.

It is to note that this might not be ideal, as it also means that the smaller dataset sizes will have a stronger KL term from the beginning, but we leave the full fine-tuning for future work. In the same vein, variational auto-encoders can be finicky to train correctly, so while the hyperparameters presented in Table 2 hold true for most experiments, we sometimes had to fine-tune it a bit further, mostly by changing the latent size.

In addition to the four generative DA algorithms mentioned in Section 4, we compare ourselves to CBERT[4], a data augmentation algorithm that showed good results on sentence classification (Wu et al., 2019). CBERT predicts masked words in

better performance to VAE-Sep. We leave the exploration of this phenomenon for future work.

[2] https://www.kaggle.com/c/fake-news/overview

[3] Code is available at https://github.com/smolPixel/DACOLING2022

[4] https://github.com/1024er/CBERT_aug

a sentence using BERT (Devlin et al., 2019). To ensure class coherence, the label is prepended to the sentence so that BERT learns to generate words of the right class. Finally, we use BERT from the jiant toolkit (Pruksachatkun et al., 2020) as our classifier.

Table 3 shows the average accuracy over the four datasets for each algorithm and starting sizes, while doubling the number of sentences[5]. Standard deviations over the 15 runs and all datasets range from 0.4 to 1.0, with CBERT and VAE-Sep achieving the lowest on average. We can also observe that VAE-Sep and VAE-Linked perform the best, surpassing the other algorithms by 0.3 points, on average. Linking the encoders and embeddings ultimately does not improve the performance over having completely separate systems. We posit that this is because the main advantage of VAE-Sep is that the separate decoders have little chance of generating examples of the wrong class, and therefore sharing the encoders is not advantageous as long as the system is good enough to help correctly train the decoder(s). We analyze this phenomenon further in Section 7. We show that even if the improvement is small in terms of performance, VAE-Sep and VAE-Linked bring a clear advantage over the other generative algorithms when looking at the number of erroneous sentences generated.

We also notice that CBERT and CVAE perform the second-best on average. While CBERT performs well when the dataset size is bigger, it underperforms on small datasets. CVAE on the other hand performs well on small dataset sizes, but badly on larger ones, similarly to the VAE-Par. Finally, we observe that the larger the initial dataset size, the harder it is to get an augmentation of performance, as noticed in (Dai and Adel, 2020).[6]

In Figure 4, we show the performances of our algorithms on the individual datasets. We see that while VAE-Sep globally outperforms other algorithms, it sometimes has more difficulties, such as for the Irony or FakeNews datasets. It is possible that this is due to the nature of the data itself, where differences between the classes come more from syntactic differences than from differences in vocabulary. This implies that chances of generating from the wrong class augment, and therefore it loses a bit of its advantage.

	500	1000	1500	2000	Aver.
Baseline	81.2	83.8	85.6	86.8	84.4
CBERT	82.2	84.5	85.8	86.9	84.9
VAE-S	**83.1**	**85.0**	**86.0**	**87.0**	**85.2**
VAE-L	83.0	84.8	85.9	**87.0**	**85.2**
VAE-P	82.2	84.0	85.3	86.3	84.5
CVAE	83.0	84.6	85.6	86.5	84.9

Table 3: Average accuracy for the four starting sizes over the four datasets. VAE-S stands for VAE-Sep, VAE-L for VAE-Linked, and VAE-P for VAE-Par. By running a multivariate t-test between the baseline results and the augmentation, we found that all presented a significant difference ($p<0.05$) except for CBERT 1500, VAE-S 1500, VAE-L 1500, CVAE 1500, and VAE-S 2000. Bold results indicate the best results for each starting sizes.

In Table 4, we show examples of generated sentences, which help give a clearer picture of the output of the algorithms. We analyze the results and the limits of each algorithms in the next section.

7 Analysis

We showed that the VAE-Sep method outperformed other methods by 0.3% globally. In this section, we analyze the algorithms and posit some hypothesis as to why it works better.

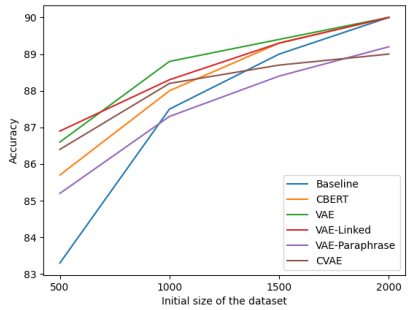
7.1 Erroneous sentences

We first take a look at the generated sentences and present the percentage which is erroneous (either has the wrong label or a human is unable to label it), as well as the percentage of sentences that are copies of genuine examples. For the former, we take the data generated for SST-2 with a starting size of 1000 (see Table 4), and manually look at 100 randomly selected examples, 50 positives and 50 negatives. For the latter, we look through all starting dataset sizes and all datasets and compute the percentage of generated sentences that are identical to a sentence in the training set (a total of 5000 sentences per algorithm). We also report various averages of length to give a better idea of the generation process, namely the length of erroneous sentences, the length of identical sentences, and the average length of generated examples.

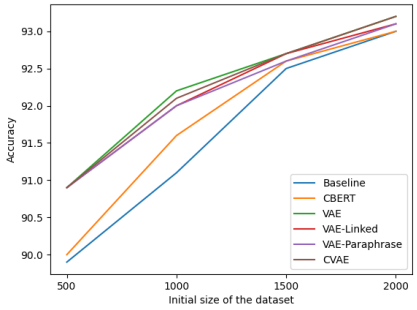
Table 5 presents the results of our analysis. We see first that VAE-Sep, VAE-Linked, and CBERT clearly produce less errors than the other meth-

[5]For reference, the maximum accuracy obtained when training with all data and no data augmentation is of 90.9% for SST-2, 95.6% for FakeNews, 69.0% for Irony, and 99.96% for Subj.

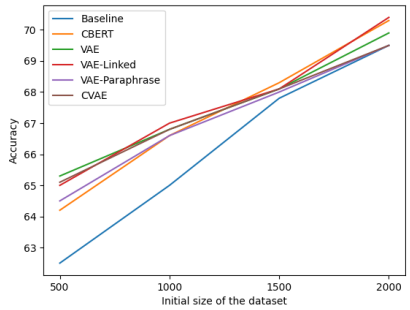
[6]We also note that even if it has a small standard deviation, the performance of CBERT is very uneven through the datasets and dataset sizes.



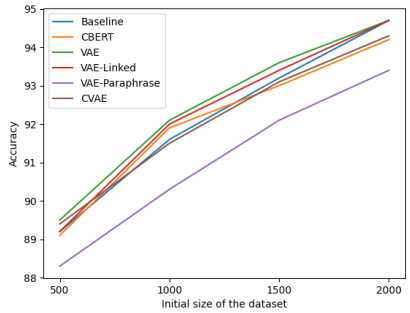
(a) SST-2



(b) FakeNews



(c) Irony



(d) Subjectivity

Figure 4: Performance of the various augmentation techniques on four dataset sizes (500, 1000, 1500, 2000) for the four datasets. Results are averaged over 15 runs.

Algo	Generated sentence	Polarity
VAE-S	a rather brilliant little cult item: a pastiche of children's entertainment, superhero comics.	Positive
	the problem with all of the characters, the characters forget their lines and no surprises.	Negative
VAE-L	it's a pleasure to watch.	Positive
	it's a frankenstein-monster of a film that doesn't know what to be.	Negative
CVAE	it is a coming-of-age, and cautionary parable, but he was a great character study.	Positive
	ultimately, the film amounts to being lectured, you's not up for the material, and offer's spiritual quest to sustain it.	Negative
VAE-P	about schmidt is nicholson's goofy, heartfelt king lear.	Positive
	the whole is a disaster, but capra are rolling in the face.	Negative
CBERT	like a tarantino movie with heart, alias betty is richly detailed, deftly executed and utterly absorbing.	Positive
	an awkwardly garish showcase that diverges from anything remotely probing or penetrating.	Negative

Table 4: Examples of generated sentences for each algorithm for the SST-2 dataset and with a dataset size of 1000.

ods, which is unsurprising due to the nature of the algorithms, and highlights the benefits of using separate decoders instead of a unique one. We note that CBERT has a very large proportion of examples that are identical to the starting examples. We find that this is a natural consequence of using a large pre-trained model. As they are trained to predict masked words, they will naturally tend to predict the correct masked words with high accuracy, and therefore often produce identi-

	VAE-S	VAE-L	CVAE	VAE-P	CBERT
Av. len	12.7	13.9	13.2	13.9	15.1
% wrong	5	6	34	27	5
Av. len	12.8	15.2	18.1	18.8	7.8
% id	1.5	6.7	1.1	2.9	43.5
Av. len	7.7	10.6	8.8	7.2	12.4

Table 5: Some statistics on the generated sentences for the five algorithms. % wrong refers to the percentage of erroneous sentences (unable to label or of the wrong class), and % id refers to the percentage of identical sentences generated by the algorithm.

Algo	Sentence
VAE-S	he's rare to be said, but it's never.
	a sleek advert youthful anomie anomie that strives for equals.
VAE-L	amazingly dopey.
	a company.
VAE-P	he doesn't take a great, and he doesn't have a very real.
	ponderous a kiss, a film is a feature film.
CVAE	an exit, while that this is that it, not entirely.
	the story is as the get-go.
CBERT	not enjoyable.
	a processed marvelous chop suey.

Table 6: Examples of erroneous sentences (pos/neg) for the SST-2 dataset with a starting dataset size of 1000.

cal examples. This also explains why their average length is much higher than for the generative algorithms, which have a tendency to produce duplicate examples when they end their generation process prematurely. CBERT, in opposition, produces duplicates as a natural consequence of its inner working. It is curious that VAE-Linked tends to produce more duplicate examples than the other generative algorithms. Most of these happens when the dataset size is small (500, 1000), and we posit that it is because using separate decoders for important characteristics (in this case, the class) makes the model more robust than a normal VAE, allowing it to have a lower reconstruction error. While this is supported by our qualitative analysis of the generated sentences of the VAE-Linked, which seems better than both generated sentences by VAE-Sep and VAE-Par, a full analysis of this phenomenon is outside the scope of this work.

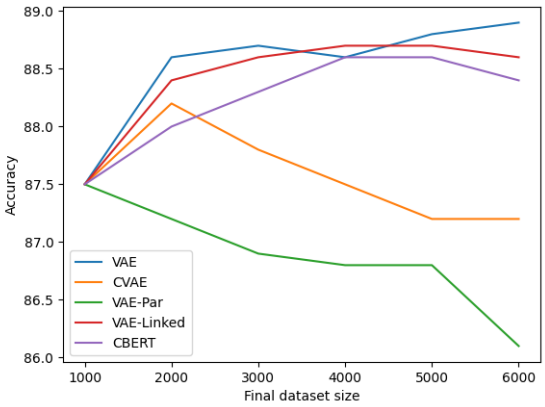


Figure 5: Accuracy vs final dataset size for a starting size of 1000 and the SST-2 task.

In Table 6, we show examples of erroneous sentences for all five algorithms. We observe that for the four generative algorithms, errors come mostly from examples that are neither positive nor negative. Rather, they are simply ill-formed. VAE-Sep and VAE-Linked create fewer of these because, as they are trained on the vocabulary of only one class, there are generally some words that pop up that allow us to determine, however tenuously, the correct class.

7.2 Number of generated sentences

Up to now, we asked the DA methods to double the size of the starting set. Here, we wonder whether generating more data would be useful, and to what extent.

Figure 5 shows for SST-2 what happens when we double, triple, etc, the size of the dataset with a starting size of 1000 sentences. The figure reveals an interesting phenomenon. None of the generative algorithms seem to benefit much from more augmentation, but CBERT does. However, the performance stops increasing at 88.6% of accuracy, which VAE-Sep reaches by simply doubling the dataset. This phenomenon is most likely directly related to the repetition of sentences produced by CBERT. In order to get the same amount of new informative sentences as VAE-Sep, the augmentation algorithm has to be applied several times. An alternative could be to introduce a filter that only accepts augmented sentences if it is not already present in the dataset, but since the focus of this study was on the VAE models and not on improving existing DA techniques, we leave this for future

work[7].

8 Implications

In this work, we show that VAE-Sep, a data augmentation algorithm that doesn't require external data, obtains a satisfying performance, even surpassing CBERT, a data augmentation algorithm using a pretrained neural network. While the VAE-Sep algorithm, at the moment, only on BERT for English sentence classification, the next step of the research is to test it on rare languages.

A large portion of NLP research focuses on English text, and even as pretrained multilingual models are allowing a good performance on multilingual data, such as XLM (CONNEAU and Lample, 2019), mBart (Liu et al., 2020), or M2M100 (Fan et al., 2021), labelled data in the target language is still needed for fine-tuning.

The VAE-Sep approach that we use in this paper does not require any form of external data, and is therefore usable for data augmentation for rare languages. We have also shown that it works well on BERT, so we expect that it would also help on multilingual pretrained transformers, most directly mBERT (Pires et al., 2019), but we leave the confirmation of this hypothesis for future work.

9 Conclusion and future work

Efficiently augmenting datasets is a central issue in machine learning. It can increase performance for all kinds of tasks, and has also been shown to help while distilling models (Kamalloo et al., 2021).

In this paper, we consider data augmentation using simple generative models based on the VAE architecture, which has the advantage that they do not need external data to work. We compare two methods used in the literature, paraphrasing and conditional generation, to another where we train one separate VAE per class. Furthermore, we compare ourselves to an established method, CBERT, which performs substitutions of random words in the sentence using a pre-trained conditionnal BERT model. We show that VAE-Sep and VAE-Linked consistently outperform other methods (CBERT, CVAE, VAE-Par), with an average gain in accuracy of 0.3%, on four binary classification tasks and four initial dataset sizes.

[7]Preliminary experiments indicate however, that it does not improve the performance of CBERT in any significant manner.

This study opens the door to interesting follow-up research. First, we contrasted generative approaches with CBERT, which showed good performance. However, due to a lack of objective comparison in the literature, we cannot guarantee CBERT delivers SOTA results. Such a comparison would help the field hone in on efficient techniques and work with a common basis. It would also be interesting to observe how VAE-Sep algorithm performs on multi-class datasets, since each VAE receives only members of one class and therefore would receive less training data on multiclass tasks.

The VAE-Sep algorithm also has possibilities of improvements. While in this study we used basic textual VAEs, there are many systems that have been developed for better VAEs that could be used. As mentioned, we take interest in VAEs because they allow data augmentation without using external data, which is useful for domains with limited data, but it would be interesting to see the performance if we allowed the usage of external data, for example by using pre-trained embeddings or even pre-trained transformers (Park and Lee, 2021; Li et al., 2020).

Finally, there is always the question of understanding why exactly data augmentation helps. While it is starting to be studied (Jha et al., 2020), we are far from fully understanding the DA process. Understanding it would not only help create stronger DA algorithms for modern deep neural networks, but furthermore help understand how these networks learn data representation.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online. Association for Computational Linguistics.

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

10–21, Berlin, Germany. Association for Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiang Dai and Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society*, 9(1):20539517221082027. Publisher: SAGE Publications Ltd.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Zubayer Islam, Mohamed Abdel-Aty, Qing Cai, and Jinghui Yuan. 2021. Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention*, 151:105950.

Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. Does Data Augmentation Improve Generalization in NLP? *arXiv:2004.15012 [cs]*. ArXiv: 2004.15012.

Ehsan Kamalloo, Mehdi Rezagholizadeh, Peyman Passban, and Ali Ghodsi. 2021. Not Far Away, Not So Close: Sample Efficient Nearest Neighbour Data Augmentation via MiniMax. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3522–3533, Online. Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*. ArXiv: 1312.6114.

Sosuke Kobayashi. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. 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.

Tomas Liesting, Flavius Frasincar, and Maria Mihaela Truşcă. 2021. Data augmentation in a hybrid approach for aspect-based sentiment analysis. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, pages 828–835, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. Place: Cambridge, MA Publisher: MIT Press.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. 2019. Controlled Text Generation for Data Augmentation in Intelligent Artificial Agents. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 90–98, Hong Kong. Association for Computational Linguistics.

Vukosi Marivate and Tshephisho Sefara. 2020. Improving Short Text Classification Through Global Augmentation Methods. In *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, pages 385–399, Cham. Springer International Publishing.

Sepideh Mesbah, Jie Yang, Robert-Jan Sips, Manuel Valle Torre, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. 2019. Training Data Augmentation for Detecting Adverse Drug Reactions in User-Generated Content. 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 2349–2359, Hong Kong, China. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Seongmin Park and Jihwa Lee. 2021. Finetuning Pretrained Transformers into Variational Autoencoders. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 29–35, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A Software Toolkit for Research on General-Purpose Text Understanding Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online. Association for Computational Linguistics.

Siyuan Qiu, Binxia Xu, Jie Zhang, Yafang Wang, Xiaoyu Shen, Gerard de Melo, Chong Long, and Xiaolong Li. 2020. EasyAug: An Automatic Textual Data Augmentation Platform for Classification Tasks. In *Companion Proceedings of the Web Conference 2020*, pages 249–252. Association for Computing Machinery, New York, NY, USA.

Juan F. Ramirez Rochac, Nian Zhang, Lara Thompson, and Timothy Oladunni. 2019. A Data Augmentation-Assisted Deep Learning Model for High Dimensional and Highly Imbalanced Hyperspectral Imaging Data. In *2019 9th International Conference on Information Science and Technology (ICIST)*, pages 362–367. ISSN: 2573-3311.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 991–1000, New York, NY, USA. Association for Computing Machinery.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2017. A Hybrid Convolutional Variational Autoencoder for Text Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 627–637, Copenhagen, Denmark. Association for Computational Linguistics.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Wang, Luis Perez, and others. 2017. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11:1–8.

Qian Wang, Fanlin Meng, and T. Breckon. 2020. Data Augmentation with norm-VAE for Unsupervised Domain Adaptation. *ArXiv*.

Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT Contextual Augmentation. In *Computational Science – ICCS 2019*, Lecture Notes in Computer Science, pages 84–95, Cham. Springer International Publishing.

Rong Xiang, Emmanuele Chersoni, Qin Lu, Chu-Ren Huang, Wenjie Li, and Yunfei Long. 2021. Lexical data augmentation for sentiment analysis. *Journal of the Association for Information Science and Technology*, 72(11):1432–1447. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24493.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2Image: Conditional Image Generation from Visual Attributes. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9908, pages 776–791. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Kang Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data Augmentation for Spoken Language Understanding via Joint Variational Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7402–7409.

Peiye Zhuang, Alexander G. Schwing, and Oluwasanmi Koyejo. 2019. FMRI Data Augmentation Via Synthesis. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1783–1787. ISSN: 1945-8452.