Class Based Variational Autoencoders

Eros Rojas
erojas20@student.ubc.ca

Bill Makwae
bmakwae@student.ubc.ca

Ryan Shar
rshar01@student.cs.ubc.ca

Abstract

Variational autoencoders (VAEs) have seen wide adoption for a variety of applications. Their ability to learn the latent space of a complex joint distribution has proven to be useful for tasks like data imputation and outlier detection. However, these models prove insufficient for "multi-modal" based applications. VAEs are unable to approximate complex, class-conditioned distributions with a high degree of accuracy. In an extreme case, this also means that traditional VAEs fail to perform well when there exists heavy class imbalances. When one label appears far more frequently than the others, standard VAEs tend to learn the distribution for the biased, more frequent label rather than the entire distribution. To remedy this, we propose the Class Based Variational Autoencoder (CBVAE). Similar to mixture models, CBVAEs learn one VAE per class and capture the distribution of features associated to each label. Additionally, CBVAEs are naturally able to perform classification tasks without modification by leveraging each learned distribution to find the most likely label. We demonstrate that CBVAEs beat traditional VAEs in both imputation and classification when class imbalance is present.

1 Introduction

Variational Auto-encoders (VAEs) have proven to be useful tools to perform outlier detection and imputation tasks. Both of these tasks are incredibly important in the real world, whether that is to detect fraudulent credit transactions (Tingfei, Guangquan, and Kuihua 2020) or to generate synthetic datasets of limited and sensitive data. However, due to their tendency to model the most dominant class labels in a dataset, they do not do as well of a job when handling heavily biased datasets. The goal of this paper is to demonstrate that, for heavily imbalanced datasets, VAEs perform better when separated into multiple models, each focusing on a specific class, rather than training a single model on the entire joint distribution. The benefits of this approach are two fold: First, using a "mixture of models" approach allows us to more accurately detect specific classes. Second, it allows for more accurate imputation of missing values.

When we split our one large VAE trained on all potential classes, into multiple VAEs for each class, we end up with a model that will have a lower ELBO loss and a tighter fit to its corresponding underlying distribution (Jr. 2018). Then by selecting the class corresponding to the VAE with the lowest reconstruction error, we can perform classification at a higher accuracy. This is especially useful in heavily biased datasets, where a regular VAE would learn the distribution of the biased label instead of the entire distribution including the minority label.

With the task of imputation, we expect to have a lower reconstruction error because the selected VAE will be specifically trained to recreate that specific class. By effectively creating a unimodal VAE, we eliminate the interaction effect that the classes have on each other leading to a more accurate imputation.

Course project for UBC CPSC 440/550, April 2024.

With this approach we propose the Class Based Variational Autoencoder (CBVAE).

2 Related Work

There are relatively few publications of VAEs being used to perform both outlier detection and data imputation. Eduardo et al. (2020) proposes the use of a Robust Variational Autoencoder which performs outlier detection for every cell within a tabular dataset instead of for each observation. To do this they trained a two-component mixture model for each feature. This yielded them very strong results with them matching or beating other state of the art outlier detection methods. In addition, not only did their model perform well at detecting outlier cells, but they also found that their approach was very strong at imputing the dirty data. Our paper aims to implement a simpler version of this approach by fitting k VAEs, one for each class instead of for each cell.

The reason as to *why* VAEs perform worse under multi modal conditions was explored in Daunhawer et al. (2021). They found that by training a singular VAE with multiple classes, the corresponding ELBO loss of the model ended up having higher upper bound and subsequently a looser approximation of the underlying joint distribution. This explanation underpins our decision to approach this problem with a mixture models technique. By separating into multiple VAEs we should end up with lower ELBO loss, and a stronger fit of the latent distribution.

Previous papers utilizing VAEs for imputation have been successful. In one of the first papers about the subject, McCoy, Kroon, and Auret (2018) found that VAEs outperformed a PCA approach while Camino, Hammerschmidt, and State (2019) found that VAEs were competitive against GANs in some situations. While these papers compared VAEs to other methods, our paper aims to compare the multi modal approach itself to a more unimodal approach.

3 Class Based Variational Autoencoders

Let $D = \{x_i\}_{i=1}^N$ denote the set of data points where each x_i is a vector of d features and $Y = \{y_i\}_{i=1}^N$ denote the set of class labels where $y_i \in \{1, ..., k\}$ is the label assigned to x_i. Like many classification tasks, we assume conditional independence given class labels such that each $P_c(x) := P(X|Y = c)$ are independent.

Motivated by mixture models, we propose Class Based Variational Auto-encoders (CBVAE). Similar to approaches that learn multiple modalities for the features (Daunhawer et al. 2021; Shi et al. 2019), our method aims to learn the modalities of each classification label. As shown in algorithm 1, we split the data by class label and train k separate VAEs, where each model is trained with only points from its corresponding class. In doing so, each VAE learns a different joint distribution $\hat{P}_c$ that approximates P_c, giving us a generative model per label.

Algorithm 1: Class Based Variational Auto-Encoders

Input : Dataset D, y, number of points in dataset N, number of classes k
Output : k VAE models

$M \leftarrow []$;
for $c \leftarrow 1$ **to** k **do**
  $D_c \leftarrow$ points in D where $y = c$;
  Train Variational Autoencoder M_c with points D_c;
  $M \leftarrow M \cup M_c$
end
return M

Since we consider each VAE to be independent of each other, we will train and find a seperate latent distribution $q_{\phi,c}$ to describe each $\hat{P}_c$. We then minimize the loss separately for each autoencoder using the standard ELBO loss listed below.

$$\text{ELBO}_{\theta,\phi}(x) = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - KL\left(q_\phi(z|x)||p_\theta(z)\right)$$

The advantage of CBVAE becomes apparent with class imbalance. Suppose we have a dataset where a large majority of labels (more than 99%) are for a single class b. A vanilla VAE trained on the joint $P(X, Y)$ will be heavily biased towards generating points with class b, implicitly learning the shape of $P(X|Y = b)$ and treating points from the other class like outliers. These vanilla VAEs will tend to have degenerate predictions, outputing only b for classification as it treats other points as outliers. For imputation, vanilla VAEs will likely use the distribution $P(X|Y = b)$ for reconstruction of missing values, even in the presence of the non-bias class.

By using CBVAE, we guarantee learning a generative model for each class regardless of the imbalance. While our learned $P(X|Y = c)$ for $c \neq b$ (where c is under-represented) may not be an excellent approximation due to little data, this is preferable to missing the points altogether and having no approximation for $P(X|Y = c)$ at all. This heavy class imbalance is present in many real world classification tasks like fraud detection, rare disease detection (Banerjee et al. 2023), and manufacturing defect detection (Rotter et al. 2023).

3.1 CBVAE for Imputation

Classically, VAEs have been been used for imputation tasks, where a subset of features have missing values (McCoy, Kroon, and Auret 2018). For our simulated imputation, we will denote missing values in our dataset with a randomly generated value from $N(0, 1)$. In order to impute these values, we will train a CBVAE over each class (using the masked data) and subsequently replace the missing values with their reconstructed counterparts. The loss function will minimize the reconstruction error between the reconstructed masked output and true (unmasked) input, thus learning a replacement for the missing values. Since every individual input to the CBVAE has a random set of m masked features, the corresponding output to the CBVAE will will replace the initial m random features with the resulting m reconstructed features.

To use CBVAE for imputation, we assume that the class label is not missing and the corresponding CBVAE model for that class (subsection 3.2) has methods for finding classes. Each VAE then learns to reconstruct the joint distribution P_c given its masked values. By using the specific $\hat{P}_c$ to impute our missing values, we are able to use a more specific distribution P_c compared to the vanilla $P(X, Y)$. This gives more precise imputations per class and allows for lower reconstruction loss across all classes.

3.2 CBVAE for Classification

A collection of CBVAE models can be used for classification by comparing each learned $\hat{P}_c$. Since we are assuming that P_c distributions are independent, the true label of $\tilde{x}$ is $\text{argmax}_c\{P_c(\tilde{x})\}$. Therefore, we use our learned approximations $\hat{P}_c$ to classify a test point $\tilde{x}$ with $\text{argmax}_c\{\hat{P}_c(\tilde{x})\}$. In order to compare $\hat{P}_c$, we compare the reconstruction loss of each model on $\tilde{x}$. If $\tilde{x}$ (or a point similar to $\tilde{x}$) appears often in some class c, we expect the VAE trained for c to reconstruct $\tilde{x}$ with little loss. Similarly, we expect a large reconstruction error for $\tilde{x}$ for classes where $\tilde{x}$ does not often appear as those VAEs have not been trained on those points. So, the class corresponding to the model that produces the lowest reconstruction error is assigned to $\tilde{x}$, as seen in algorithm 2.

4 Experimental Results

From our experiments, we find that CBVAE outperforms a vanilla VAE in classification and imputation for imbalanced datasets. We used the Fraud Detection dataset where 99.8207% of samples are labeled "Not Fraud" while only 1.793% were labeled "Fraud." In each experimental environment, we trained our CBVAE autoencoders and the vanilla autoencoder with the same hyper-parameters.

4.1 CBVAE for Imputation

For our imputation experiment, we trained a VAE, $P(X, Y)$, on a representative sample of the Fraud Detection dataset (not split by class) as a baseline, as well as two CBVAE's, P_c, (one for each individual class). For both models, 14 of the 28 features were masked with random noise, each row having a random selection of 14 features. After training, each model was evaluated on an unseen testing set of 70 "Not Fraud" and 70 "Fraud" observations. The testing set was also masked with

Algorithm 2: CBVAE Classification

Input : A point $\tilde{x}$, CBVAE models M
Output : The class label

best_error $\leftarrow \infty$;
for $c \leftarrow 1$ **to** k **do**
    $x_c \leftarrow$ reconstruction of $\tilde{x}$ using M_c;
    $MSE_c \leftarrow$ Mean squared error($\tilde{x}$, x_c);
    if $MSE_c <$ *best_error* **then**
        best_error$\leftarrow MSE_c$;
        best_label$\leftarrow c$;
    end
end
return *best_label*

random noise according to the training procedure. The VAE was given the entire testing set to reconstruct, where the CBVAE's were given their respective testing points to reconstruct, aggregating the loss over both classes.

When computing the overall loss of the testing set, it was found that the per-feature variance of the reconstruction error was substantially larger for the baseline VAE when compared to the CBVAE's. More specifically, the variance in reconstruction for every feature was higher on average for the baseline VAE as compared to the CBVAE's.

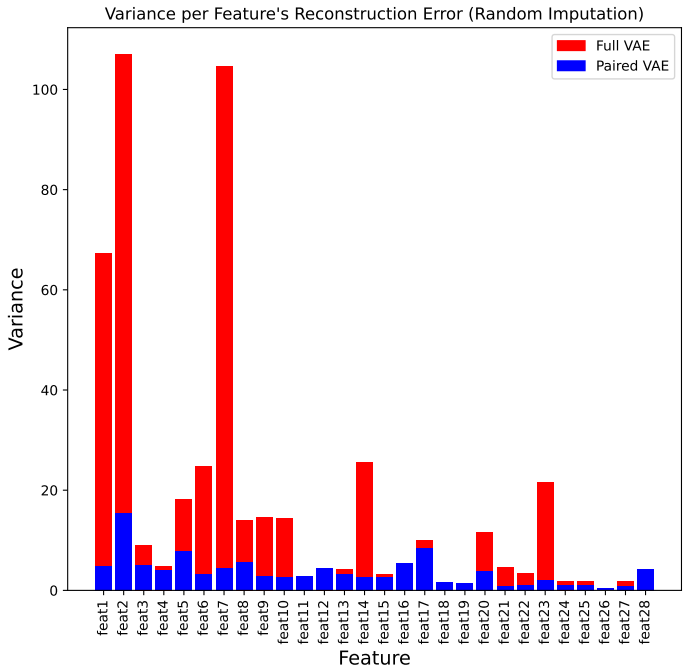

Figure 1: Variance per feature, colour coded for both models. The values of variance are overlapped on one another, thus the heights of the bars correspond do the variance per feature.

As shown above, the variance of the CBVAE's reconstruction error is consistently lower than the variance for the baseline VAE, for virtually all features. These results are consistent with subsection 3.1. Since we are able to generate independent representations for $\hat{P}_c$, we are able to create better latent space approximations for P_c and thus better data reconstructions for every class, particularly for the "Fraud" case. In the baseline VAE, all "Fraud" testing points achieved extremely

variable reconstruction losses due to the fact that the model was was unable to learn a meaningful representation of these points due to the extreme imbalance.

4.2 CBVAE for Classification

In our experiments, we found that the vanilla VAE only predicted a single label regardless of the X that was inputed. As we predicted, the vanilla VAE treated "fraud" cases as outliers due to their infrequency. Our CBVAE was trained with "random masking" on two features and on five features. The CBVAE classifier was able to distinguish between fraud point some of the time as seen in Figure 2 and Figure 3.

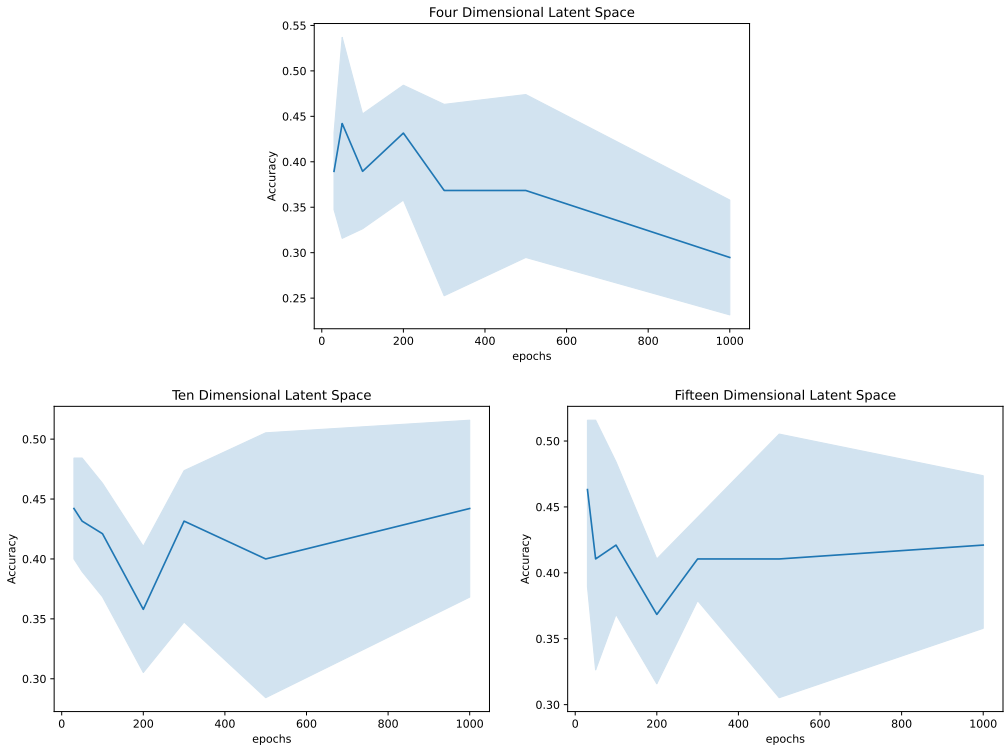

Figure 2: Loss per epoch when asking two features at a time. The plots are separated by model architecture, with different latent dimensions. The model with four dimensional latents (top) is overfitting while models with ten dimensional (bottom left) and fifteen dimension (bottom right) latents continue to learn.

The classification accuracy contains high variance due to the limited data available for the unbiased class. The VAE was unable to fully generalize the distribution with the amount of data which affected the performance based on seed.

5 Discussion

We have demonstrated that CBVAEs are able to outperform traditional VAEs in imputation tasks when given extremely imbalanced datasets. Additionally, we showed that CBVAEs have some effectiveness at classification tasks by comparing each learned $\hat{P}_c$ to find the most likely label for a given point.

Due to lack of computation, we were unable to search over many architectures and parameters. Future works can extend our model with higher dimensional data and more complex VAE models like MVAE (Shi et al. 2019) and MoPoE-VAE (Sutter, Daunhawer, and Vogt 2021). Additionally, future works can extend CBVAEs using the idea of weighted VAEs from Daunhawer et al. (2021). Weighted CBVAEs can have learned responsibilities for each classifier, predictions can be made using a weighted approximation of reconstruction error.

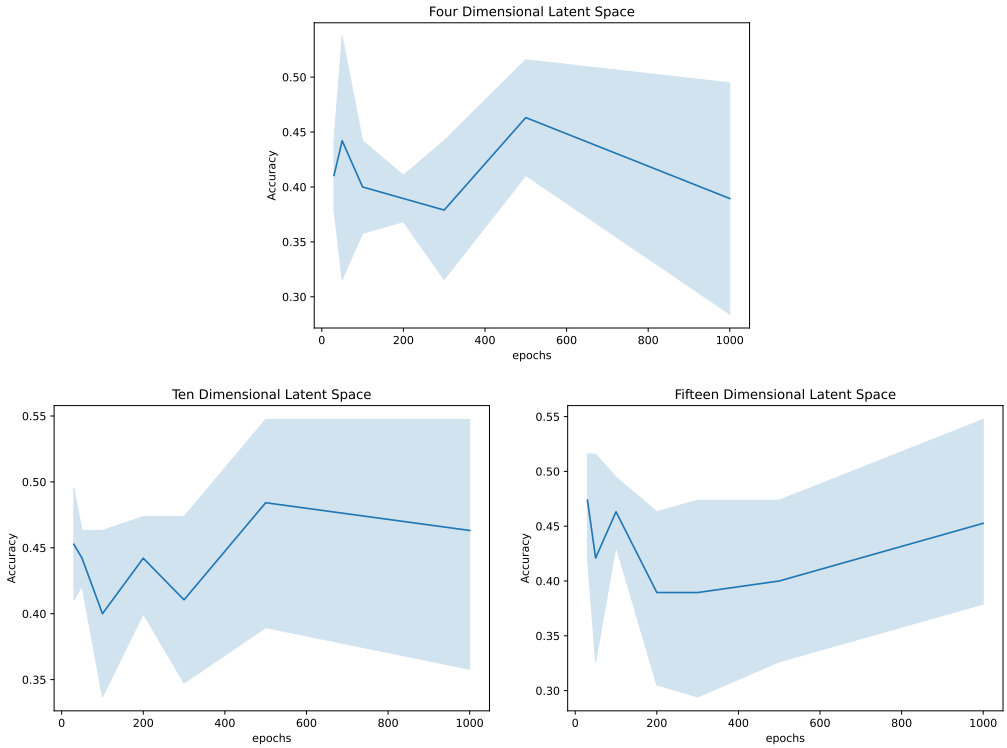

Figure 3: Loss per epoch when masking five features at a time. The plots are separated by architecture different latent dimensions. The model with four dimensional latent (top) overfits while the ten dimensional latents (bottom left) and fifteen dimension (bottom right) continue to learn.

References

Banerjee, Jineta, Jaclyn N Taroni, Robert J Allaway, Deepashree Venkatesh Prasad, Justin Guinney, and Casey Greene (June 2023). "Machine learning in rare disease." en. *Nat. Methods* 20.6.

Camino, Ramiro Daniel, Christian A. Hammerschmidt, and Radu State (2019). "Improving Missing Data Imputation with Deep Generative Models." *CoRR* abs/1902.10666. arXiv: 1902.10666.

Daunhawer, Imant, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E. Vogt (2021). "On the Limitations of Multimodal VAEs." *CoRR* abs/2110.04121. arXiv: 2110.04121. URL: https://arxiv.org/abs/2110.04121.

Eduardo, Simão, Alfredo Nazábal, Christopher K. I. Williams, and Charles Sutton (2020). *Robust Variational Autoencoders for Outlier Detection and Repair of Mixed-Type Data*. arXiv: 1907.06671 [cs.LG].

Jr., Ally Salim (2018). "Synthetic Patient Generation: A Deep Learning Approach Using Variational Autoencoders." *CoRR* abs/1808.06444. arXiv: 1808.06444.

McCoy, John T., Steve Kroon, and Lidia Auret (2018). "Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit." *IFAC-PapersOnLine* 51.21, pp. 141–146. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2018.09.406.

Rotter, Dominik, Florian Liebgott, Daniel Kessler, Annika Liebgott, and Bin Yang (2023). "Machine Learning-Based Identification of Root Causes for Defective Units in Manufacturing Processes." Cham: Springer International Publishing, pp. 168–178.

Shi, Yuge, Siddharth N, Brooks Paige, and Philip Torr (2019). "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models." *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Sutter, Thomas M., Imant Daunhawer, and Julia E. Vogt (2021). "Generalized Multimodal ELBO." *CoRR* abs/2105.02470. arXiv: 2105.02470. URL: https://arxiv.org/abs/2105.02470.

Tingfei, Huang, Cheng Guangquan, and Huang Kuihua (2020). "Using Variational Auto Encoding in Credit Card Fraud Detection." *IEEE Access* 8, pp. 149841–149853. DOI: 10.1109/ACCESS.2020.3015600.