You’re your own best teacher: A Self-Supervised Learning Approach For Expressive Representations

Johan Vik Mathisen

June 1, 2024

In this section we present the results from both the experiments on stage 1 and 2, while addressing the research questions.

To summarize NC-VQVAE is able to capture the conditional distribution of the data better than naive VQVAE for a wide variety of datasets. For datasets with few training samples, our model can be prone to overfitting. We see our model as a step in the right direction, but further development is needed to ensure better intraclass diversity, possibly through a more refined sampling procedure.

Key takeaways: We are able to simultaneously reconstruct well and significantly improve downstream classification accuracy, which is very interesting from a representation learning perspective. We improve both IS, FID and CAS for most datasets, indicating that the conditional distribution is better captured, as well as the synthetic data being closer to the ground truth. Additionally we see some differences in Barlow Twins and VIbCReg when it comes to sample diversity.

NC-VQVAE is better able to mimic the training data. When data is abundant, then our model better captures the entire distribution, while covering

Some of the issues of TimeVQVAE are still highly relevant, such as the difficulty in modelling data with sharp differences in modularity, such as TwoPatterns and ElectricDevices.

0.1 Stage 1

In this section we present the results of the tokenization model, in particular the reconstruction loss and the downstream classification accuracy. We address research question 1 and 2, if the proposed NC-VQVAE is able to reconstruct on par with the naive VQVAE and if the learned latent representations are more expressive, in the sense that they simultaneously improve the downstream classification accuracy. We see that some configuration of NC-VQVAE is the top performer on the majority of datasets for both metrics, and provides significant increase in probe accuracy.

0.1.1 Reconstruction

We present top 1 and mean reconstruction loss across the four runs in table **??** and table **??** respectively.

Mean validation reconstruction error

Dataset	Baseline	SSL Method					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.217	0.127	0.134	**0.108**	0.173	0.169	0.203
ElectricDevices	**0.041**	0.067	0.044	0.049	0.105	0.042	0.049
StarLightCurves	**0.032**	0.042	0.069	0.071	0.052	0.050	0.068
Wafer	0.044	0.037	0.048	0.049	**0.035**	0.042	0.039
ECG5000	**0.048**	0.083	0.170	0.104	0.093	0.205	0.064
TwoPatterns	0.197	0.201	**0.184**	0.230	0.214	0.186	0.207
UWaveGestureLibraryAll	0.190	**0.172**	0.190	0.245	0.189	0.178	0.237
FordB	0.150	0.115	0.122	0.123	**0.114**	0.121	0.142
ShapesAll	**0.045**	0.056	0.066	0.102	0.064	0.069	0.073
SonyAIBORobotSurface1	0.402	0.509	0.494	0.491	**0.360**	0.363	0.418
SonyAIBORobotSurface2	0.623	0.622	0.618	0.640	0.487	**0.454**	0.589
Symbols	0.110	0.143	0.134	0.173	0.078	**0.067**	0.105
Mallat	0.066	0.081	0.091	0.096	0.066	0.067	**0.060**

Table 1: Mean validation reconstruction error across all 13 datasets. Results are averaged over four runs.

Top 1 validation reconstruction error

Dataset	Baseline	SSL Method					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.158	0.108	0.111	**0.087**	0.130	0.134	0.113
ElectricDevices	0.036	0.060	0.034	0.043	0.092	**0.031**	0.045
StarLightCurves	**0.026**	0.037	0.057	0.055	0.043	0.048	0.065
Wafer	0.038	0.031	0.045	0.043	**0.027**	0.031	0.038
ECG5000	**0.044**	0.069	0.156	0.084	0.080	0.181	0.056
TwoPatterns	0.181	0.184	**0.169**	0.208	0.200	0.172	0.185
UWaveGestureLibraryAll	0.159	**0.145**	0.167	0.201	0.155	0.169	0.233
FordB	0.117	0.094	0.090	0.103	**0.082**	0.094	0.102
ShapesAll	**0.035**	0.043	0.046	0.092	0.061	0.063	0.067
SonyAIBORobotSurface1	0.381	0.473	0.472	0.465	0.329	**0.328**	0.408
SonyAIBORobotSurface2	0.513	0.577	0.536	0.588	0.444	**0.414**	0.470
Symbols	0.088	0.111	0.122	0.150	0.062	**0.059**	0.090
Mallat	0.061	0.075	0.076	0.088	0.059	0.059	**0.057**

Table 2: Top 1 validation reconstruction error across all 13 datasets. Lowest value of the four runs for each model is selected.

From the tables we see that NC-VQVAE reconstructs on par with the baseline model, and that some configuration outperforms the naive VQVAE on mean reconstruction loss for 9 out of 13 datasets. In figure **??** observe that the difference in reconstruction loss is small for most datasets, both across SSL methods and

augmentations. The use of gaussian augmentation introduces less of a regularizing effect compared to the two others, with the exception of Slice and shuffle on ECG5000. These results show that the introduction of a non contrastive loss does not hurt the reconstruction capabilities of our model compared to naive VQVAE.

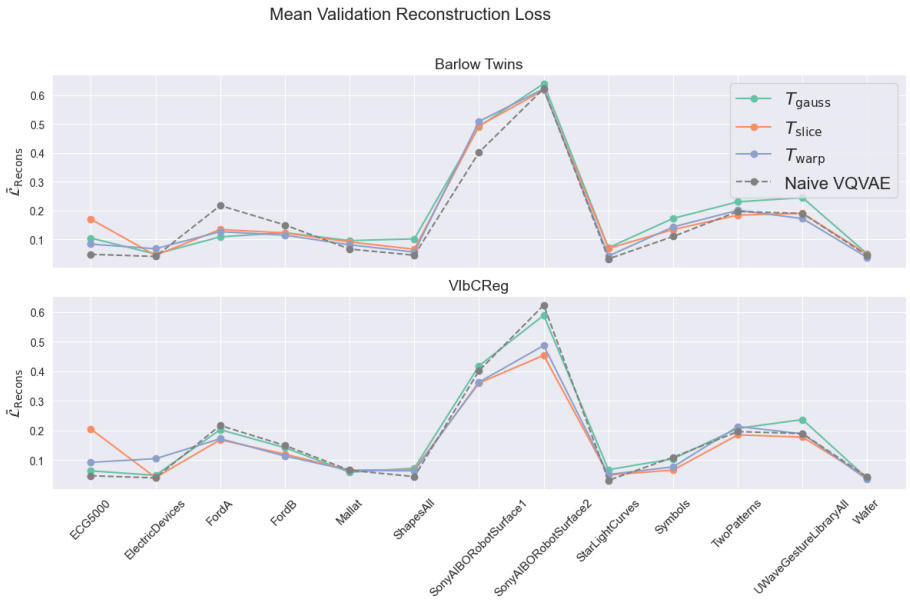

Figure 1: Mean validation reconstruction loss for the two models, compared to naive VQVAE

The right configuration for NC-VQVAE acts as a regularizer, in figure **??** we see how the validation reconstruction loss developes on FordA during training.

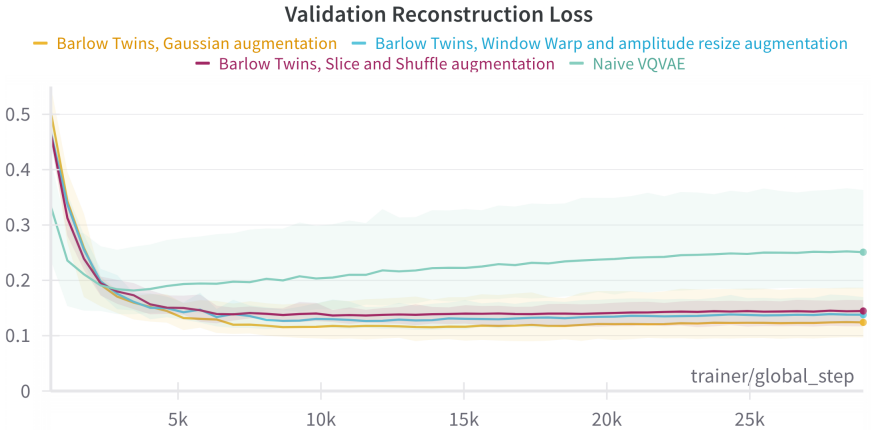

Figure 2: Development of the validation reconstruction loss for FordA during training. Averaged across all four runs.

0.1.2 Classification

We present the mean and max downstream classification accuracy in table **??** and **??** respectively.

Mean linear probe accuracy

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
FordA	0.70	0.74	0.83	0.84	**0.91**	**0.89**	0.80	0.83	0.80	0.74	0.87	0.86	0.76	0.78
ElectricDevices	0.35	0.41	0.35	**0.44**	0.38	0.41	**0.40**	0.42	0.33	0.38	0.36	0.39	0.39	0.43
StarLightCurves	0.87	0.89	0.93	0.93	**0.94**	**0.94**	0.88	0.88	0.92	**0.94**	0.91	0.93	0.89	0.89
Wafer	0.93	0.89	0.96	**0.94**	0.96	**0.94**	0.96	0.93	**0.97**	0.94	0.96	0.92	**0.97**	0.92
ECG5000	0.80	0.83	0.85	0.81	**0.88**	0.84	0.86	**0.84**	0.86	0.82	**0.88**	**0.84**	0.84	0.82
TwoPatterns	0.34	0.53	**0.69**	**0.91**	0.66	0.82	0.47	0.71	0.64	0.90	0.68	0.80	0.55	0.72
UWaveGestureLibraryAll	0.31	0.40	**0.62**	0.70	0.56	0.63	0.40	0.54	**0.62**	**0.73**	0.55	0.66	0.44	0.55
FordB	0.58	0.60	0.64	0.67	0.74	0.76	0.64	0.68	0.63	0.64	0.70	0.70	0.61	0.64
ShapesAll	0.29	0.30	0.49	0.55	0.53	**0.60**	0.40	0.48	0.48	0.56	**0.54**	**0.60**	0.40	0.46
SonyAIBORobotSurface1	0.56	0.68	0.54	0.70	**0.61**	**0.74**	0.53	0.70	0.48	**0.74**	0.58	0.71	0.54	0.69
SonyAIBORobotSurface2	**0.81**	**0.86**	0.77	0.79	0.80	0.80	0.80	0.81	0.77	0.85	0.80	0.85	0.80	0.85
Symbols	0.50	0.60	**0.59**	0.60	0.50	**0.66**	**0.59**	**0.66**	0.45	0.61	0.42	0.62	0.43	0.63
Mallat	0.63	0.77	0.72	0.81	0.76	0.83	0.68	0.78	**0.79**	**0.87**	0.77	0.85	0.69	**0.86**

Table 3: Summary of mean linear probe accuracy by SSL Method and Augmentation. Average across 4 seeds. Best result for KNN and SVM are highlighted in bold.

Top 1 linear probe accuracy

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
FordA	0.75	0.78	0.84	0.88	**0.93**	**0.92**	0.85	0.87	0.81	0.77	0.88	0.90	0.86	0.85
ElectricDevices	0.35	0.43	0.36	0.45	0.39	0.43	**0.45**	**0.46**	0.34	0.42	0.39	0.42	0.42	0.45
StarLightCurves	0.89	0.91	0.94	0.95	**0.96**	**0.96**	0.90	0.91	0.95	0.95	0.93	0.95	0.90	0.90
Wafer	0.94	0.89	**0.97**	**0.95**	**0.97**	**0.95**	**0.97**	0.93	**0.97**	**0.95**	**0.97**	**0.95**	**0.97**	0.94
ECG5000	0.83	0.84	0.88	0.86	**0.90**	**0.88**	**0.90**	**0.88**	0.88	0.85	0.89	0.86	0.86	0.85
TwoPatterns	0.37	0.62	**0.75**	**0.96**	0.68	0.85	0.55	0.75	0.70	0.92	0.71	0.81	0.63	0.76
UWaveGestureLibraryAll	0.34	0.43	**0.67**	0.74	0.60	0.67	0.43	0.54	**0.67**	**0.76**	0.58	0.67	0.48	0.58
FordB	0.60	0.63	0.67	0.71	**0.76**	**0.80**	0.69	0.74	0.67	0.65	0.74	0.77	0.63	0.68
ShapesAll	0.33	0.34	0.53	0.59	**0.59**	**0.65**	0.44	0.50	0.50	0.56	0.57	0.63	0.44	0.48
SonyAIBORobotSurface1	0.67	**0.80**	0.61	0.77	0.76	**0.80**	0.60	0.74	0.51	0.79	0.63	0.75	0.63	0.75
SonyAIBORobotSurface2	**0.84**	**0.89**	0.80	0.86	0.82	0.84	0.83	0.82	0.81	0.88	0.81	0.88	0.83	0.87
Symbols	0.56	0.66	0.65	0.69	0.55	0.73	**0.64**	**0.71**	0.51	0.65	0.45	0.67	0.46	0.69
Mallat	0.54	0.88	0.57	0.87	0.74	0.89	0.66	0.80	0.74	**0.92**	0.72	0.88	0.62	**0.90**

Table 4: Summary of max linear probe accuracy by SSL Method and Augmentation. Maximum value across 4 seeds. Best result for KNN and SVM are highlighted in bold.

We observe a significant improvement in probe accuracy with NC-VQVAE, compared to naive VQVAE. Some configuration is best on 12 out of 13 datasets, while the one where our model falls short, the difference is one percent for both SVM and KNN. The largest differences are seen on FordA, FordB, Mallat, ShapesAll, TwoPatterns and UWaveGestureLibraryAll.

Even though the specific choice of augmentation has a large impact, it is interesting that on the majority of datasets, all choices result in improved probe accuracy.

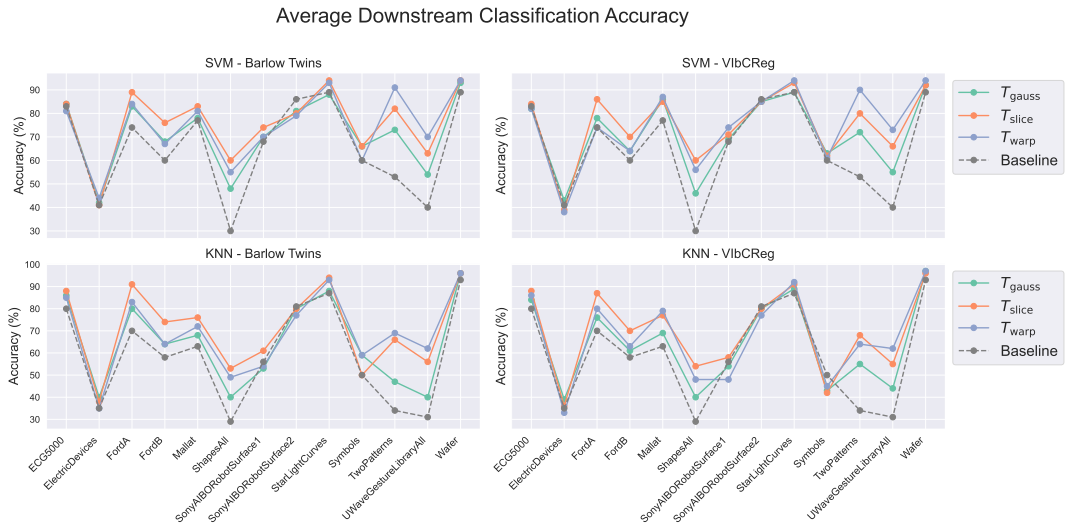

Figure 3: Mean probe accuracies.

These results show that NC-VQVAE is able to reconstruct on par with naive VQVAE, and in some cases improve the reconstruction loss, while significantly improving the probe accuracy for the discrete latent representations for most datasets. The NC-VQVAE produces representations that separates classes more effectively, and could in turn encode more class specific information.

0.1.3 Losses

TODO: How does the minimization of diffrerent losses influence FID/IS?

We investigate some trends in the development of different loss terms in this section. To summarize, using VIbCReg results in more easily minimizable losses compared to Barlow Twins, and the Gaussian augmentation results in significantly easier minimization of the SSL loss as well as reducing the VQ loss.

In figure **??** we observe the typical pattern of the SSL loss during training.

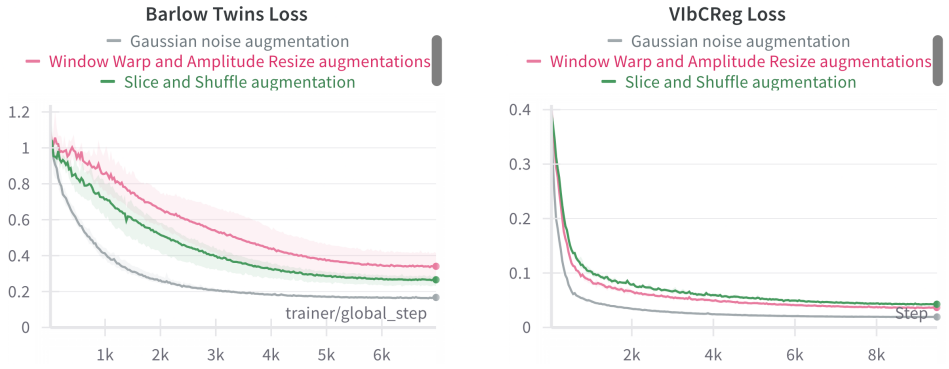

Figure 4: SSL loss during training on UWaveGestureLibraryAll. Averaged across four runs. The pattern shown here is typical across datasets. The Gaussian noise augmentation leads to an easier minimization, while the two others are relatively similar with the occasional dataset where one or the other is minimized the most. In general the VIbCReg loss is minimized faster than Barlow Twins.

We see that the Gaussian augmentation results in a SSL loss which is easier to minimize, which might be attributed to the fact that it affects the samples in a more predictable way. We too see that the VIbCReg loss decreases more rapidly than the Barlow Twins loss. Both observations are too seen across datasets. Previously in figure **??** we have seen that the gaussian augmentation often resulted in lower probe accuracy than the other two. In figure **??** we see that on the datasets with a significant increase in probe accuracy, the augmentations that result in a more challenging SSL loss typically has higher downstream classification accuracy. We also see that for a specific augmentation, the pattern is rather flat, indicating that the particular augmentation plays the most prominent role in probe accuracy. The SSL loss for a specific augmentation varies very little compared to probe accuracy. This could be due to the fact that augmentations such as Slice and Shuffle often alter the semantics of the sample and could make the SSL loss push latent representation to more unpopulated areas of the latent space, making it easier to classify, but harder to minimize.

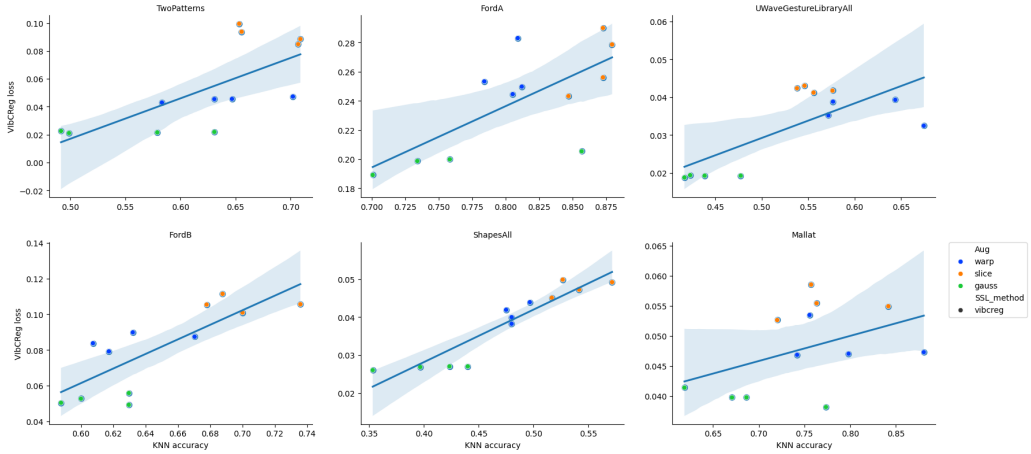

Figure 5: KNN accuracy plotted against VIbCReg loss. Each point correspond to a single run of the model. Similar tendency is shown for Barlow Twins.

The training reconstruction losses are heavily minimized, both across models and augmentations. The only consistent noticeable difference is the augmented reconstruction loss, where models using Slice and Shuffle have a slightly higher loss. The differences in VQ loss for the different models is mainly due to the codebook, where we too observe that VIbCReg minimizes more effectively than Barlow Twins, and again that gaussian augmentations results in the hardest minimization followed by Window Warp and then Slice and Shuffle. Both VIbCReg and Barlow Twins with Gaussian augmentation routinely perform on par with naive VQVAE in terms of VQ loss during training. The minimization of the codebook loss indicates that the encoder is properly aligned with the discrete latent codes. We hypothesize that when the SSL loss is not properly minimized, the encoder must adjust its weights more throughout training which keeps the encoder outputs and the discrete codes from aligning completely.

TODO: What effect might this have on stage 2??

0.1.4 Visual inspection

After training the stage 1 model, we map the training and test data to discrete latent representations using the learned encoder and codebook. The representations are then global average pooled before being mapped onto the plane by TSNE showed in Figure **??**.

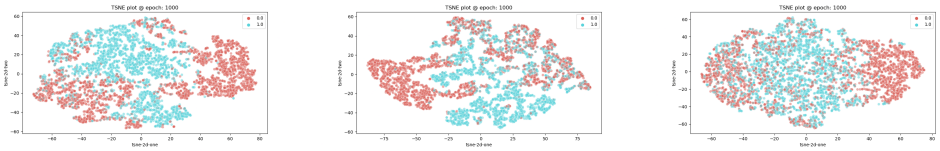

Figure 6: TSNE plots of FordA. Barlow (left) and VIbCReg (center) with Slice and Shuffle, naive VQVAE (right). Best performing model in terms of KNN accuracy is chosen.

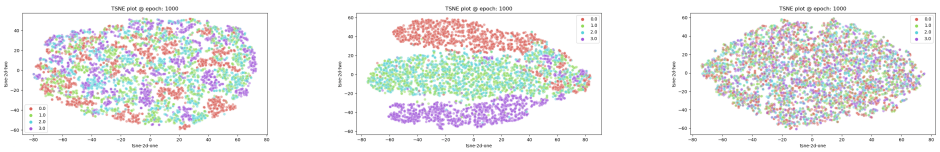

Figure 7: TSNE plot of discrete latent representations from VIbCReg with Slice and Shuffle (left), Barlow Twins with Window Warp and Amplitude Resize (center) and naive VQVAE (right). Dataset is TwoPatterns. The latent space is significantly more structured with NC-VQVAE.

From these visualizations it is evident that representations learned using NC-VQVAE are more structured than those of the naive VQVAE. Similar samples, typically with the same label, are clustered closer together in latent space.

0.2 Stage 2

For datasets with very few samples, or very few per class, the generative scores must be taken with a grain of salt. Both the classifier, and the evaluation metrics is dependent on a certain number of samples to be considered reliable. We rather look more closely on the visual inspection for these.

0.2.1 Generative quality

The generative quality of or models are evaluated according to FID, IS and CAS. We present the top 1 results in table **??**, and the mean score across the four runs in table **??**. From the tables we see that our model produces better IS score for 12 out of 13 datates, and better FID for 10 out of 13.

Top 1 FID and IS

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS ↑	FID↓	IS↑
FordA	2.59	1.30	1.93	**1.51**	2.13	1.48	1.80	**1.51**	2.83	1.38	2.50	1.43	**1.66**	1.41
ElectricDevices	12.05	3.97	11.82	4.20	**8.91**	4.07	9.89	3.86	12.38	**4.23**	11.08	3.94	13.96	3.71
StarLightCurves	**0.74**	1.99	0.89	**2.43**	1.50	2.36	0.75	2.39	0.92	2.39	0.85	**2.40**	0.79	2.26
Wafer	5.27	**1.39**	3.31	1.29	3.82	**1.26**	2.77	1.35	3.33	1.29	3.60	1.30	**2.52**	1.34
ECG5000	1.56	2.01	2.43	2.02	2.27	2.00	2.15	2.02	2.15	**2.03**	2.21	2.00	**1.52**	2.02
TwoPatterns	3.63	2.47	3.59	2.65	2.74	2.73	**2.24**	2.70	3.45	2.64	2.90	2.70	**2.19**	**2.77**
UWaveGestureLibraryAll	8.16	2.24	6.45	2.94	**6.26**	**3.13**	7.31	2.79	6.52	2.99	6.33	3.06	7.09	2.79
FordB	2.92	1.52	2.10	1.52	2.44	1.61	1.93	**1.67**	1.76	1.65	2.12	1.64	**1.66**	1.52
ShapesAll	**21.35**	4.32	35.89	**5.22**	29.61	5.16	27.91	4.83	30.03	4.95	31.59	4.92	27.20	4.94
SonyAIBORobotSurface1	18.21	1.27	26.20	1.32	28.90	1.28	21.63	1.32	21.98	1.36	25.20	1.38	**15.73**	**1.55**
SonyAIBORobotSurface2	3.85	1.69	2.50	1.82	3.34	1.79	**0.82**	1.82	2.61	1.81	2.75	1.83	1.24	**1.84**
Symbols	8.50	2.43	5.86	3.20	7.39	2.82	**4.25**	**3.50**	6.78	3.39	7.21	3.23	8.21	3.30
Mallat	**1.31**	3.41	2.01	3.67	2.24	3.72	1.85	3.66	1.87	3.34	2.30	3.05	**1.31**	**3.92**

Table 5: Summary of FID and IS scores by SSL Method and Augmentation. Best achieved results are highlighted in bold

Mean FID and IS

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS ↑	FID↓	IS↑
FordA	5.15	1.16	2.59	1.41	2.36	**1.45**	**2.28**	**1.45**	3.01	1.34	2.90	1.41	3.73	1.29
ElectricDevices	13.48	3.75	16.51	3.95	**10.20**	3.93	11.54	3.75	13.99	**4.17**	11.82	3.85	15.20	3.55
StarLightCurves	**1.01**	1.93	1.29	2.35	1.91	2.32	1.08	2.25	1.07	2.35	1.19	**2.36**	1.05	2.22
Wafer	5.72	**1.33**	3.70	1.25	4.20	1.24	2.85	1.31	3.67	1.26	3.86	1.26	**2.84**	1.31
ECG5000	**1.62**	1.94	2.61	**2.00**	2.56	1.98	2.47	**2.00**	2.60	1.99	2.39	**2.00**	1.76	1.99
TwoPatterns	4.04	2.41	4.00	2.54	2.96	2.66	**2.44**	**2.67**	4.05	2.56	3.15	2.66	2.62	**2.67**
UWaveGestureLibraryAll	8.48	2.13	6.77	2.86	6.64	2.96	7.35	2.73	6.80	2.91	**6.49**	**2.99**	7.34	2.72
FordB	4.05	1.28	2.66	1.48	3.49	1.50	2.88	**1.52**	**2.49**	1.48	3.07	1.51	3.04	1.31
ShapesAll	**27.64**	4.22	38.22	5.07	32.54	**5.04**	32.25	4.56	36.59	4.72	35.79	4.76	31.56	4.71
SonyAIBORobotSurface1	23.71	1.20	30.65	1.22	31.97	1.21	25.29	1.28	26.11	1.32	28.20	1.32	**18.61**	**1.44**
SonyAIBORobotSurface2	5.42	1.62	3.35	1.77	4.41	1.74	**1.78**	**1.81**	4.43	1.74	3.32	1.79	2.36	1.79
Symbols	13.62	1.99	9.78	2.92	9.78	2.67	8.61	3.14	8.84	3.20	9.74	3.03	**8.58**	**3.24**
Mallat	2.09	3.01	2.54	3.29	3.68	2.94	2.12	3.53	2.11	3.18	2.40	2.96	**1.65**	**3.72**

Table 6: Summary of FID and IS scores by SSL Method and Augmentation. Best mean achieved FID and IS are highlighted in bold

In figure **??** we get a better overview of the results, and observe that both Barlow Twins and VIbCReg produces better samples than the naive VQVAE in terms of FID and IS. Additionally we see that the use of gaussian augmentation results in the largest improvements for most datasets. The high IS scores indicate that NC-VQVAE captures the conditional distributions better than naive VQVAE in many datasets. This will be explored further in section **??**. The improved FID scores indicates that the synthetic samples more closely resemble the test data. The moderate decrease in FID, compared to the increase in IS, could indicate that the generated samples does not generalize too well to the test data. The discrete latent representations from NC-VQVAE provides more information regarding the classes, as we saw from the improved downstream classification accuracy in stage 1. This additional class specific information seems to assist the prior learning in capturing class conditional distributions.

Figure 8: Mean FID and IS scores for Barlow Twins and VIbCReg VQVAE. FID is plotted on a log scale because of the large difference in values across datasets.

It is worth mentioning that the FID and IS score is calculated using the SupervisedFCN, which is also trained on the UCR Archive. Thus, the FID and IS scores could have a bias toward samples that mimic the training data.

0.2.2 Class conditional sampling

TODO: How do we calculate the CAS? How many samples etc.

We present the mean CAS for all models across datasets in table **??**.

Mean CAS

Dataset	**Baseline**	**SSL Method**					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.864	0.884	**0.902**	0.878	0.864	0.895	0.870
ElectricDevices	0.614	0.588	0.607	0.599	**0.618**	0.610	0.594
StarLightCurves	0.960	0.953	0.955	**0.965**	0.962	0.954	0.964
Wafer	0.976	0.977	0.978	0.968	0.979	0.976	**0.984**
ECG5000	0.866	0.881	0.863	0.880	0.877	0.892	**0.910**
TwoPatterns	0.808	0.770	0.788	**0.847**	0.715	0.781	0.846
UWaveGestureLibraryAll	0.333	0.300	0.367	0.313	0.360	**0.401**	0.383
FordB	0.725	0.748	**0.756**	0.741	0.750	0.738	0.750
ShapesAll	0.361	0.344	0.329	**0.420**	0.379	0.367	0.404
SonyAIBORobotSurface1	0.975	0.933	0.957	0.979	0.982	0.976	**0.985**
SonyAIBORobotSurface2	0.929	0.956	0.951	0.969	0.960	**0.970**	0.964
Symbols	0.956	0.929	0.930	0.930	0.969	**0.974**	0.963
Mallat	0.471	0.642	0.563	0.661	0.827	0.876	**0.908**

Table 7: Mean CAS score across datasets. Results averaged across four runs.

We see that some configuration of NC-VQVAE outperforms the naive VQVAE on all datasets, as well as VIbCReg with gaussian augmentation outperforming the baseline on 12 out of 13, where the one dataset where it falls short its within one percent. In general we observe that NC-VQVAE performs well across all datasets, and in particular with gaussian augmentation. The dataset where we see the most dramatic increase is Mallat, with an improvement of 0.437. This particular case will be investigated in section **??**.

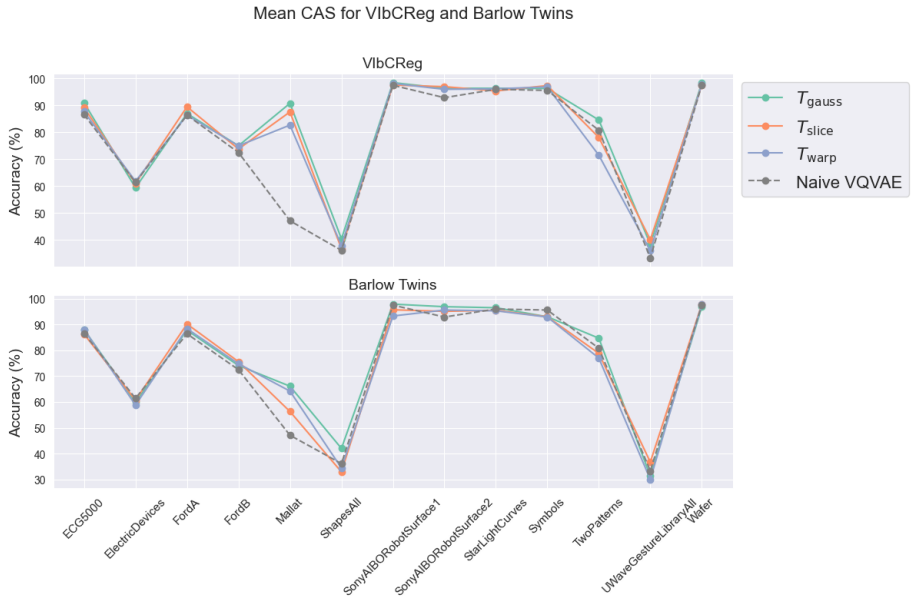

Figure 9: Mean CAS across all datasets.

0.2.3 Prior loss

Mention that during experiments with our stage 2 modification, embed / fine-tune, we observed that the val prior loss with our modification was higher, but with similar shape as without. If we had time and computational resources to re-run the experiments, then we would omit the stage 2 modification. The FID/IS in our main experiments are in many cases better than baseline VQVAE, despite higher val prior loss.

Naive VQVAE outperforms NC-VQVAE in terms of validation prior loss across datasets. There is the occasional dataset where a model with gaussian augmentation performs equally well. The minimization of the validation prior loss does therefore not correspond to improved synthetic samples in general.

0.2.4 Token usage

We say memorization/overfitting when the selected probabilities are mainly >0.9.

wafer: Barlow and VIbCReg are more certain of tokens than naive. Often at time T=3 the main proportion of selected samples have probability >0.9. Barlow to a greater degree than VIbCReg.

ShapesAll: Barlow a bit more uncertain than VIbCReg. For a seed they both collapse and basically sample tokens with probability 1 from T=1.

Sony2: Barlow is much more certain earlier for several models. Both are significantly more certain than naive.

Mallat: All models overfit quite hard. VIbCReg and Barlow has some more variability, vibcreg best of SSL.

FordB: more healthy distributions. VIbCReg a bit more certain, in a good way i think.

ECG5000: For several models barlow and vib overfits hard. Naive has healty distributions, mostly.

TwoPatterns: Naive consistently uncertain. VIbCReg and Barlow develops similarly as T increases, and looks very good. This is a prime example of what i consider good.

UWave: VIbCReg and Barlow has several cases of severe overfitting, . Naive looks healthy.

Symbols: VIbCReg has significantly more diversity than barlow. Still overfits quite a bit. Naive overfits in some cases, but generally healthier.

ElectricDevices: Similar behavior. After T=4 almost certain.

StarLightCurves: Overfitting in some cases. Otherwise a more healty distribution than naive.

Sony1: Barlow overfits more than VIbCReg . Both have some quite extreme cases. Warp produces the best distributions.

FordA: One model each with some overfitting (both gaussian).

Note: Does overfitting etc, happen mostly for Gauss? It is the case for StarLightCurves.

Include something on the differences in sampling/token usage between naive VQVAE and NC-VQVAE. NC-VQVAE has a tendency to be more certain of tokens selected. For small datasets such as Mallat, the certainty is close to 1 for most sampled tokens.

TODO: Investigate this further. Compare/relate the selected probability histograms with token usage histograms / perplexity

Would be interesting to investigate different values for T in maskgit iterative sampling.

Higher masking ratios during training etc.

0.2.5 Visual inspection

Typically naive vqvae has more trouble with capturing the global consistency of the samples when samples are scarce and diverse, as seen on ShapesAll and Symbols. In contrast, our method will tend to overfit in these cases. The overfitting issue is most prominent in the class conditional distributions.

For the most simple shapes, NC-VQVAE is better at maintaining the little variability that is present.

Simple patterns, such as sinusoids, are very easily captured (Symbols/ShapesAll). Sharp changes in modularity and frequency are much harder(TwoPatterns).

Datasets or classes with very few samples might be mimicked/overfitted.

Different samples sizes, how easy patterns are etc should be considered when setting nr of epochs, and T in maskgit (lower for simple patterns).

There are only minor differences seen in the generated samples form the models trained with different augmentations, especially when sample size is low.

ECG5000

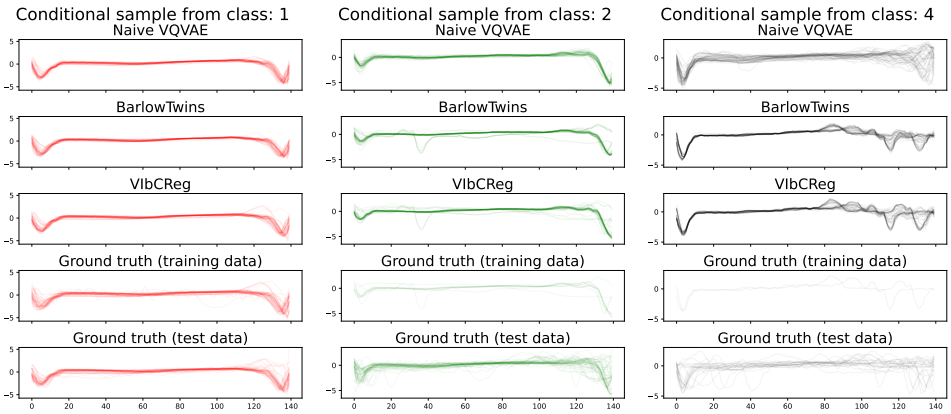

Figure 10: Class conditional distribution for some selected classes of ECG5000. Barlow and VIbCReg both trained with window warp and amplitude resize augmentations.

Mallat

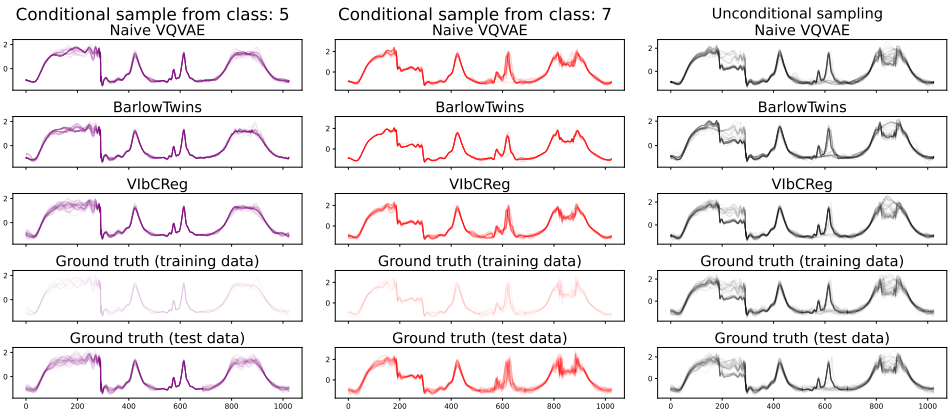

Figure 11: Class conditional distribution for some selected classes of Mallat, in addition to unconditional samples. Barlow and VIbCReg both trained with gaussian augmentation.

Symbols

Naive VQVAE: Does not capture the entire underlying distribution, some classes are not represented/not recognizable. Global consistency for the sinusoids are poor, particularly towards the end.
VIbCReg: Good mode coverage, but underrepresents the sinusoids or lacks diversity in each class.

Barlow Twins: windowwarp: little variability in sinusoids, could it be that the ssl loss makes these too close in latent space?

Does the high IS scores correlate with good mode covarage? For symbols our model cover the modes much better than naive. But produces many very similar samples. Does this have something to do with the selected token histograms. Seems like our models select tokens with higher probability, sometimes much higher!

IS has a flaw in that it does not take intraclass diversity into account. Thus a model which generates the mode at each class will get a high IS score. Thus it can give high scores to models that overfit.

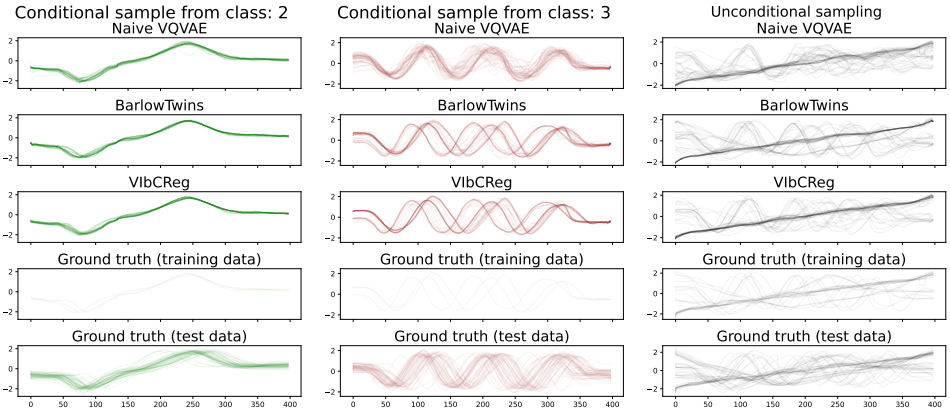

Figure 12: Class conditional distribution for some selected classes of Symbols. Barlow and VIbCReg both trained with gaussian augmentation.

ShapesALL

A LOT BETTER class conditional sampling!

Generated vs real.

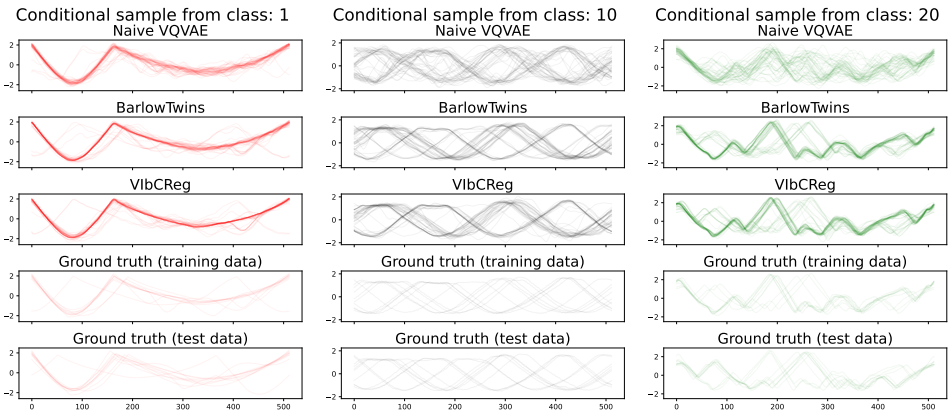

Figure 13: Class conditional distribution for some selected classes of ShapesAll. Barlow and VIbCReg both trained with gaussian augmentation.

UWaveGestureLibraryAll

As before in TimeVQVAE, sharp changes in modularity is challenging to model.

There is evidence that gaussian augmentation results in less diverse samples, which most likely is a result of ??. Most evident in Barlow Twins. VIbCReg is able to maintain variability to a greater degree (See class 1). Most to least diverse: warp, slice, gauss for VIbCReg, slice, warp, gauss for Barlow Twins. VIbCReg does best in terms of capturing the flat spots after the "break", though not great.

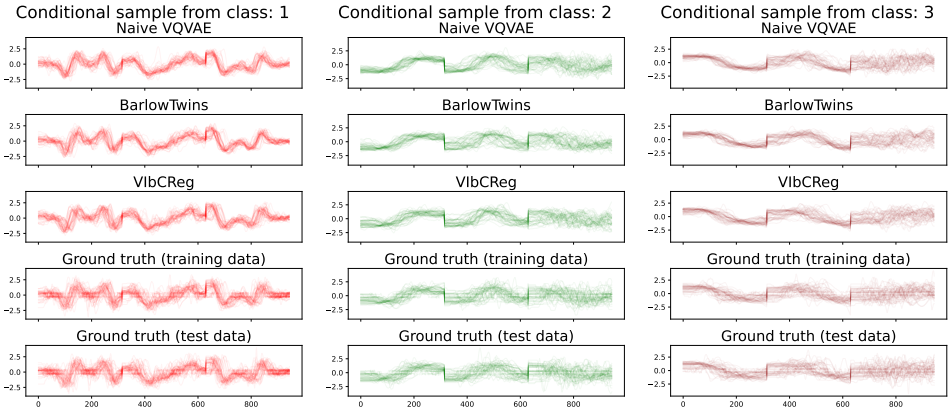

Figure 14: Class conditional distribution for some selected classes of UWaveGestureLibraryAll. Barlow and VIbCReg both trained with window warp and amplitude resize augmentations.

0.3 The influence of stage 1 on stage 2

The best performing datasets in terms of probe accuracies: "FordA", "FordB", "Mallat", "ShapesAll", "TwoPatterns", "UWaveGestureLibraryAll"

Relationship between reconstruction in stage 1 and FID/IS/CAS: Does better reconstruction capabilities in stage 1 improve the generative model?

Relationship between probes in stage 1 and FID/IS/CAS: Does better probe accuracies (class separation) in stage 1 improve the generative model?

How does the best performing models from stage 1 transfer to stage 2?

Look at FordA, FordB, Mallat, ShapesALL, TwoPatterns and UWaveGestureLibraryAll. The datasets where probe accuracies are good compared to baseline. Slice is aug with best performance overall on these datasets.

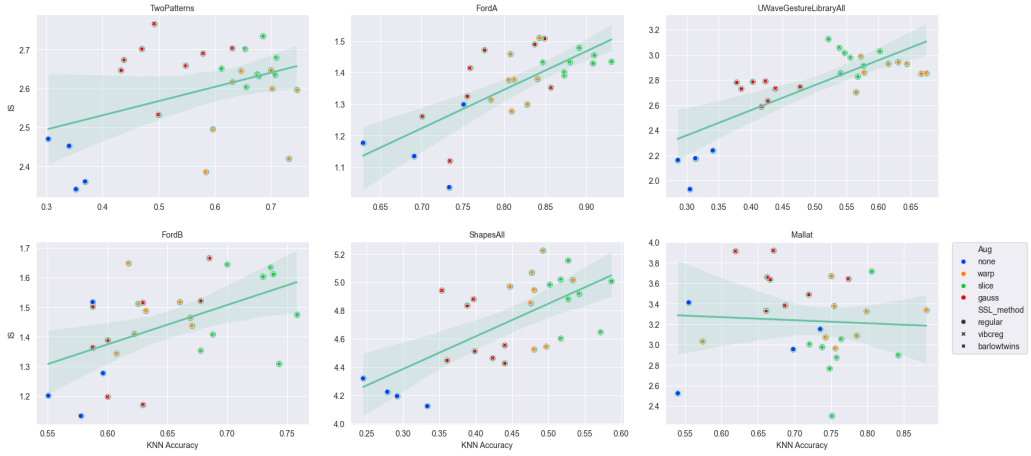

Figure 15: KNN plotted against Inception Score on the subset of datasets with significant improvement in probe accuracy.

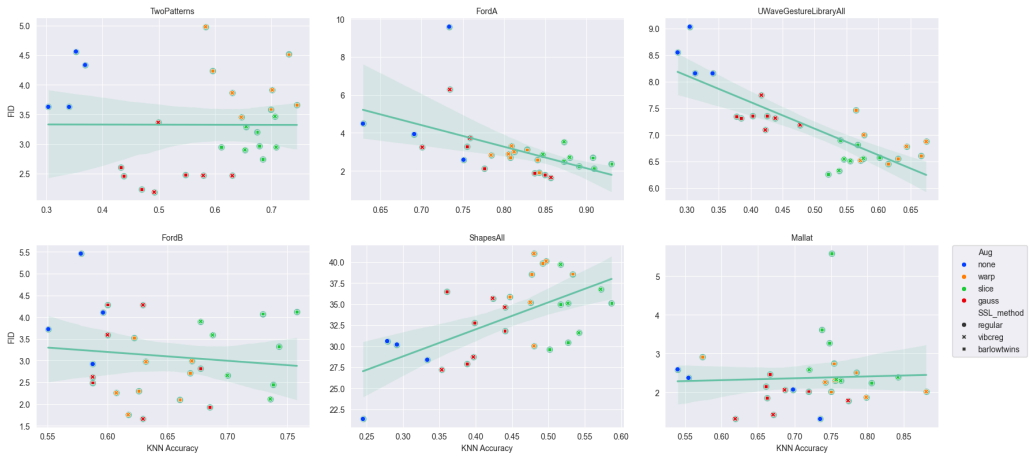

Figure 16: KNN plotted against Fréchet Inception Distance on the subset of datasets with significant improvement in probe accuracy.

In Figure **??** and **??**, we see the relationship between KNN accuracy and FID/IS on the subsets where the probe accuracy is substantially improved with NC-VQVAE. The corresponding plots with SVM accuracy show similar trends. From Figure **??** we see a trend, with higher probe accuracy correlating with higher IS. Upon closer inspection, we see a clear pattern of the prominent effect of augmentations. For each specific augmentation, the correlation between KNN and IS is close to 0.

Particularly interesting is UWaveGestureLibraryAll.

0.4 Differences in Barlow Twins and VIbCReg

VIbCReg seems to keep the variablility in the conditional distribution a bit better than Barlow Twins. Can it be attributed to the variance term in VIbCReg?

TSNE and PCA of Mallat.

0.4.1 Overfitting problem

0.4.2 Thoughts

Better inception score and CAS of our models indicate that the class separability learned in latent space makes the conditional distributions more distinct easier to classify. The FID is variable, but in many cases better, which indicated that the generative distributions are closer to the ground truth.

Gaussian noise aug seems to result in a lot easier the BT/VIbCReg loss to minimize.
Slice and shuffle is harder to minimize, but could seem to push representations for different classes further apart resulting in better linear probes.

Talk about the difficulty/ease in minimizing the SSL loss for the different augmentations. Does this affect linear probes / reconstruction / FID / IS / Prior loss

For datasets of smaller size with classes of different characteristics (a clear distributional difference in visual inspection [Sony2 and Symbols]) NC-VQVAE seems to perform better both in terms of FID and IS.

The biases introduced by augmentations in stage 1 seems to be included in the generated samples to some degree. In particular datasets with high frequency components, when applying Gaussian noise (easier to spot), has substantially better FID score.

Is there correlation between CAS and linear probe accuracy??

Temporal vs frequency influence of augmentations. We compress only along temporal axis in the encoder. Could this be a reason for Gaussian artifacts in generation and not slice?

0.5 Discussion

Even though there are issues, we believe our model is a step in the right direction. The representations learned by NC-VQVAE are more expressive than naive VQVAE, demonstrating that we can optimize more than one objective, without sacrificing the reconstruction capability.

The representations enable easier learning of the semantics of the conditional distributions, to such a degree that one has to take measures not to overfit.

The representations makes it easier to capture the gobal consistency of the samples, which tha naive VQVAE has large issues with. This without the HF-LF split.

The added flexibility of NC-VQVAE, with possibility of choosing dataset specific augmentations, can in some applications be beneficial.

0.6 Further work

[**morningstar2024augmentations**] suggest that focus on augmentations is of great importance. The hunt for good augmentations in the time series domain is ongoing and should probably get more attention.
HF-LF split - augmentations tailored for HF and LF, as they often have quite different characteristics.
Wavelet transform to improve HF-LF split.

Further optimize the relationship between aug recon loss and choice of augmentations.

Improving on the stage 2 learning to better handle the expressive representations, and be able to create more diverse samples. Higher masking ratio during training, lower value for T etc.

The differences in Barlow and VIbCReg indicate that further optimization of the SSL method/pretext task for generative performance is possible and could be an interesting extension of this project.

Investigation of the attention maps. Could one find if there is a HF direction, translation direction etc. similarly to how one in NLP can find gender direction, nationality direction etc?