Ice Core Data Analysis

> get_participants()

'Fabian Depenau Bjørnholt Jacobsen', 'Povl Filip Sonne-Frederiksen', 'Magnus Guldbæk Hansen', 'Marcus Melhedegaard Thomsen'



Overview

- Introduction
- Data
 - MNIST
 - Artificial dataset (labeled dataset from Nicolo)
 - Peruvian Ice Core Samples
- Methods
 - Classifier (Resnet18)
 - AutoEncoder (Resnet18 + Decoder)
 - Variational AutoEncoder (Resnet18 + Decoder)
- Analysis
 - Classifier
 - AutoEncoder
 - Variational AutoEncoder
- Findings
- Further Work



Introduction

The goal is to use dimensionality reduction to identify "interesting" objects within the Peruvian ice core dataset.

٠	٠	ø		9	٠	9	4		A	٠	3	۳	
Index = 2093	Index = 4196	index = 934	index = 2234	Index = 2399	Index = 262	index = 2993	Index = 2940	index = 770	index = 4251	Index = 3322	Index = 2032	Index = 3545	index = 3803
2			0	0			•		0		3	•	0
Index = 3344	Index = 4019	Index = 1783	index = 4286	Index = 2362	Index = 1499	index = 4426	Index = 841	Index = 1509	index = 3797	Index = 994	Index = 2566	Index = 2350	Index = 1234
0	٠		4	٠	3	3	0	٠.	3.	Þ			٠
Index = 872	Index = 3870	Index = 3697	Index = 2386	index = 810	Index = 2406	Index = 2921	Index = 4278	Index = 3555	Index = 2005	Index = 2535	Index = 1321	Index = 3490	Index = 765
6	٠			9		0	•		8	3	0	D	0
Index = 3796	Index = 2999	Index = 1660	index = 2963	Index = 479	Index = 2927	index = 3784	Index = 974	index = 802	index = 3111	Index = 129	Index = 1627	Index = 3557	index = 1777
1	0		٠	٠		1	۰	-		-		P	
Index = 3492	Index = 3953	Index = 1340	Index = 2963	Index = 3112	Index = 328	Index = 2546	Index = 1502	Index = 3757	index = 2538	Index = 4134	Index = 3112	Index = 3171	Index = 2047

Data

A look at the datasets and their structures

MNIST

Handwritten digits

- 70.000 samples
- 28x28p size
- Balanced
- Preprocessed



MNIST

Handwritten digits

- 70.000 samples
- 28x28p size
- Balanced
- Preprocessed



MNIST - Dataset distribution

Artificial

Image samples

- 145.242 samples
- Resized to 128x128p
- 34 scalars
- 7 Labels
- Unbalanced



Artificial

Image samples

- 145.242 samples
- Resized to 128x128p
- 34 scalars
- 7 Labels
- Unbalanced

	Particle ID	Area (ABD)	Area (filled)	Aspect Ratio	Biovolume (Cylinder)	 Fiber Straightness	Filter Score	Geodesic Aspect Ratio	Geodesic Length	Geodesic Thickness	
0	3733	79.16	79.97	0.55	446.21	 98.0	0	0.32	17.59	3.64	
1	55.2.8	9.55	9.55	0.67	47.64	 12	0	1.0	3.93	3.93	
2	27	711.31	716.33	0.71	13121.26	 1.05	0	0.6	36.07	21.52	
3	2780	54.6	54.6	0.5	293.64	 0.55	0	0.37	13.57	5.19	
4	3510	41.53	43.93	0.77	209.93	 0.6.9	0	0.34	13.33	4.48	
5	456.0	49.15	49.15	0.6	227.46	 0.79	0	0.3	54.92	4.41	
6	251	492.14	3453.44	0.5	30433.77	 1.01	0	0.41	61.07	25.19	
7	3730	20.15	20.15	0.5	302.9	 0.85	0	0.56	7.43	4.2	
8	35.9	34.46	30.44	0.42	154.61	 0.89	0	0.29	13.16	3.87	
9	1041	215	2.16	0.55	12.67	 1.05	0	1.0	2.53	2.53	
30	338	3010.61	3043.2	0.56	57890.21	 0.67	0	0.17	135.95	23.28	
11	925	953.0	953.0	0.24	11772.3	 0.99	0	0.21	69.55	14.67	
12	36.3.0	6.57	6.57	0.59	32.51	 126	0	1.0	3.46	3.46	
13	1024	24.32	25.07	0.53	95.4	 0.79	0	0.29	11.43	3.26	
54	4421	34.97	36.12	0.54	178.42	 0.92	0	0.4	11.23	4.5	
15	24910	11286.34	12293.55	0.29	217302.8	 0.5	0	0.04	568.22	22.07	
36	9695	1356.39	1357.79	0.94	22830.78	 0.63	0	0.25	71.32	20.19	
17	611	125.32	125.32	0.29	695.85	 1.06	0	0.23	25.77	2.00	
15	114	406.39	505.6	0.59	9016.04	 124	0	10	22.56	22.56	
29	7737	1339.99	2474.94	0.46	16961.02	 0.67	0	0.12	112.62	13.56	
20	76	10.36	10.36	0.71	55.07	 112	0	1.0	4.12	4.12	
21	2504	27.5	27.5	0.44	70.27	 0.6	0	0.12	18.07	2.22	
22	82	494.23	501.66	0.87	9878.65	 1.22	0	1.0	23.26	23.26	
23	952	3541.71	3354.44	0.8	100845.08	 0.71	0	0.26	124.11	32.16	
4	2002	3.06	3.06	433	2.0%	 121	0	10	323	323	
2	436	301.41	301.41	0.32	11550.00	 105	0	10	30.47	10.14	
77	4700	10.84	10.84	0.01	113.07	 113	0	10	5.20	5.74	
27	725.1	21.55	21.55	0.00	174.05	 174	0	10	3.47	547	
	2558	10.19	10.19	0.55	11.00	11		10	407	4.07	
10	75.7	0.4	0.4	0.19	111	0.95		10	119	119	
31	5305	815.77	515.77	9.75	13897.45	0.57	0	0.46	43.73	20.12	
32	359.2	24.01	24.01	0.24	27.0	 0.77	0	0.03	37.65	0.96	
33	336	1587.91	1907.26	0.63	34429.57	 0.34	0	0.04	218.66	9.17	
34	17161	401.0	402.07	0.51	4071.36	 0.55	0	0.3	35.6	11.59	
35	96.7	21.24	21.24	0.5	115.7	 114	0	1.0	5.28	5.28	
36	2569	1076.36	1053.77	0.55	21755.35	 0.89	0	0.5	48.35	25.94	
37	22.7	35.1	35.1	0.55	369.29	 0.53	0	0.39	11.25	4.37	
35	40	405.67	459.97	0.9	6501.59	 0.82	0	0.48	33.64	36.04	
30	92.7	1139.05	1143.45	0.81	15045.25	 0.67	0	0.2	77.54	15.72	
40	1382	91.9.66	942.76	0.82	15861.03	 0.79	0	0.39	50.65	29.97	
41	15995	80.1	80.1	0.6	355.55	 0.72	0	0.24	20.6	4.9	
42	4574	21.67	21.67	0.63	115.23	 0.97	0	0.63	7.28	4.55	1
43	5382	958.75	958.75	0.89	20170.92	 0.92	0	0.61	41.14	24.98	1
44	19.9	476.10	504.41	0.55	4743.64	 0.56	0	0.21	50.98	10.88	
45	1024	39.91	41.27	0.57	203.65	 0.52	0	0.37	12.34	4.58	1
46	4215	47.71	45.29	0.2	135.1	 0.5	0	0.11	24.38	2.66	1
47	3207	7.0	7.D	0.84	30.9.5	 1.04	0	1.0	14	3.4	
45	32.47	602.16	604.12	0.0	4836.51	 0.57	0	0.13	70.58	9.34	
42	1371	394.59	394.89	0.55	5590.32	 119	0	0.5	27.16	16.19	

Artificial

Image samples

- 145.242 samples
- Resized to 128x128p size
- 34 scalars
- 7 Labels
- Unbalanced



TRAIN - Dataset distribution

Peruvian Ice Core Samples

Image samples

- 102.763 samples
- Resized to 128x128p
- 34 scalars
- Probably extremely unbalanced with a lot being dust



Peruvian Ice Core Samples

Image samples

- 102.763 samples
- Resized to 128x128p size
- 34 scalars
- Probably extremely unbalanced with a lot being dust

	Bergin	Postdate 13	Arriva VARE1	Anna Pillanti	Angent Patie	 Pillan-Xaman	Demokratis Report Platter	Desident's Length	Gradesis Teldeses	Investiger Film	
	8710	82	8.12	2.01	834	 	10	102	430	\$7,31,1,1,40,25,0000Lif	
1	11934	47	8.17	36.17	14		14	120	1.25	\$7,28,4,3,540,25x,00000.w	
	80.0		8.73	8.71	107		544	3.43	4.000	\$2,210,1040,00,00000AF	
	362.74	83	8.75	8.75	105		246	27.05	1.02	\$7,31,4,3,140,25,38080.w	
4	101.2	411	1.0	1.0	14		11	417	4.87	07,22,5,2,640,255,08086.W	
	100.71		4.000	4.85	141			127	146	WARDON CONTRACTOR	
	W11.0	58	8.00	26.00	194		14	143	143	07,25,4,4,940,20x,000805.w	
,							6.07	1.41		OT 25 & A DIRECTOR DISTORT	
	m1m		8.1	2.1	087		645	834	4.85	(17.34.4.2.540.20x.000806.wr	
,	2105	10	8.9	8.71	175		14	145	145	07,21,4,3,540,20x,000802.w	
	*****		8.11	8.11			6.00	8.1	1.00	177.38.2.2.1.40.30x.00000.47	
	36.26	80	8.11	8.11	1.7		14	6.07	6.07	(17, 21, 3, 1, 1, 40, 20x, 080802.w/	
	#P13	84	32	12	047		67	189	225	07,25,7,4,140,254,080805.W	
	817.76		1.0	8.21	671		647	8.0	427	177.32 A 1 (40) 20x (10) 001 of	
	8845	3.11	8.0	10.09	175			115	853	07,21,3,3,640,25x,0000644F	
	80.95	916	2.17	19.30	071		14	4.00	4.000	\$25,38,3,3,3,46,3%,080805.W	
	w11.1	-	1.01	3.44	545			445	445	177.28.4.8.040.20x.08080.wr	
	8043	20	34.05	2.0	1N			1.71	121	07,28,4,3,640,20,00000.w	
	#379		37.58	37.58	047		14	4.03	4.83	07.35.7.3.540.20x.08080.w	
	math	87	679	629	041		14	362	382	\$7,32,5,2,540,25,08086.w	
	107.54		8.71	8.21	061		587	8.0	196	\$7,38,4,3,640,26,08088.W	
*	20.0	843	8.00	8.00	049		1.00		100	\$2531.0.1.140.0%.00006.W	
	8.172	201	2.01	2.01	042		640	122	363	07.31.0.1.40.26.08086.w	
		,	80.5.1	87.89	0.000		11	8.14	8.14	07,22,1,2,040,25,08080.W	
			W.12	80.12			14	657	887	\$7.32.3.3.540.26.00006.w	
	2047	270	12.13	12.13	0.72		ш	196	136	07,28,33,040,25,00000.w	
	26.5	83	31.4	36.4	547		629	*1	363	\$25,22,3,2,4,6,00,00,00000.AF	
	201.00		20.00	20.00	641		605	11.12	218	\$7,37,3,1,140,26,38086.w	
	4115		2.0	2.00	042		63	141	283	07,21,2,2,140,25x,00000.w	
	A.11	#1	83.55	82.00	147		642	31.45	24	07,31,0,1,140,05,00000.w	
	2015	87	81.25	25.38	041		014	2.44	2.71	\$7,21,5,4,140,25x,00000.w	
	m47%	26	162	162	05.3	 	613	1.1	1.05	\$2,35,4,3,540,30x,00000.w	
	862.76	-	8.54	8.54	847	 	605	34.00	204	\$2,34,4,3,540,356,00000.w	
	BC 06	35	2.11	8.10	044		627	2.11	144	07,21,4,3,140,20x,00000.w	
-	890	837	17.04	12.44	5m		645	3.42	143	\$27,21,4,3,540,30x,00000.w	
-	B1924	6	4.00	4.00		 	14	184	184	\$7,31,4,3,540,26,00006.w	
	80.95	-	x.m	¥.38	1.7		643	854	1.00	\$7,28,4,1,140,254,00000.W	
*	1773.4	818	8.0	4.11	041		1.00	32.54	4.67	\$2,34,3,3,640,85,00000.w	
-	main		8.43	8.43	1.74	 	625	8.02	4.00	\$7,31,0,1,40,26,00082.w	
	811		10.00	10.00	114	 	605	1.11	1.79	07,22,3,4,040,85s,000805.W	
	96.776	10.14	110	839		 	14	167	167	07,34,3,3,140,35,00000.w	
4	96177	24	8.16	¥.41	1M		ч	614	614	\$7,31,4,3,140,254,00000.w	
	W176		36.13	26.13	184		м.	487	4.87	\$7,35,4,0+80,85,08088.W	
-	21100		8.14	8.14			и	875	873	\$7,34,3,3,144,25,36006.w	
	80%	84	8.11	2.11	853	 	61	8.0	186	07,28,4,3,140,25s,00000.w	
-	8815	2.01			967		м	1.76	1.76	\$7,31,3,3,540,356,00006.w	
-	3106		8.11	8.11	041		ч	440	440	\$7,31,4,3,40,25,38086.w	
	8313	41	2.0	2.0	044		043	684	4.36	07,21,4,1,140,254,00005.w	
-	371.0	100	3.43	3.43	145		642	8.0	101	\$7.34.3.4.340,85.00000.M	
-	1	1									

Peruvian Ice Core Samples

Image samples

- 102.763 samples
- Resized to 128x128p size
- 34 scalars
- Probably extremely unbalanced with a lot being dust



Presumed class distribution in the Peruvian Dataset

We remove all that the classifier characterizes as dust from the Peruvian set and look at the rest. We include only images above 15x15 pixels in size

Methods

How did we engage with the data

Classifier – Resnet18 on MNIST

• See if classification can be used to find unknow categories by mapping the n-1 layer in the network



Classifier – Resnet18 on Artificial and Peruvian Datasets

• See if classification can be used to find unknow categories by mapping the n-1 layer in the network



AutoEncoder

• See if the bias can be reduced by encoding a latent space



Variational AutoEncoder

• Regularizing the latent space to reduce overfitting



Example of latent space for AE vs VAE



Analysis

Our results

Classifier – MNIST

Mapping the second to last layer in ResNet18 where [2, 6] have been left out of the MNIST Train Dataset



AutoEncoder – MNIST

Mapping the latent space of aM AutoEncoder trained on the MNIST Train Dataset



Variational AutoEncoder – MNIST

Mapping the latent space of a Variational AutoEncoder trained on the MNIST Train Dataset



CNN-Classifier on Artificial Dataset



CNN-Classifier on Peruvian Dataset



AutoEncoders on Peruvian Data



Decoder: VAE-ResNet18-PyTorch/model.py at master · julianstastny/VAE-ResNet18-PyTorch (github.com)



AutoEncoder plain



AutoEncoder with mean correction and no Dust



AutoEncoder plain latent space 128

Variational AutoEncoders on Peruvian Dataset



planar-flow-pytorch/vae-pf.py at master · abdulfatir/planar-flow-pytorch (github.com)

Difficulties during training

- Overfitting on initial parameters?
- KL divergence for some events in the Peru set explodes before training has even started and never seem to drop. Most likely overfitting encoder from classifier on the artificial set. And yes! Problem solved by reducing epochs on pretrained encoder.



Difficulties during training – mode collapse?

Input images

KLD was about 2 vs about 40 in "good" trainings indicating information depleted latent space.

Difficulties with overfitting



Fitted too much on labeled data?

Plotting scalars



KMeans Clustering





KMeans for n=7















Overlaying labels from CNN-Classifier



DBSCAN

• DBSCAN on the AE/VAE tensors allows for partial clustering



Artificial Dataset (Prediction)

Artificial Dataset (Truth)

Peruvian Dataset







DBSCAN on the AE/VAE tensors allows for partial clustering

Consistency between mappings

Plotting the loss


CNN classifier - Area 1





CNN classifier - Area 2





Between area 1 and 2



CNN classifier - Area 3





CNN classifier - Area 4



VAE, planar flow Area 1



VAE, planar flow area 2





Findings

Our results

20 Images

Main categories



Outliers



Separate area of high-resolution images similar area #5 (Outlier)



Low circularity image surrounded by ones with high circularity. (Outlier)



Low circularity images surrounded by ones with high circularity. (Outlier) Unusually deform









Further work

- Try implementing auto encoders with added scalar output.
- Use permutation importance on scalar parameters in classifier to reduce the number of parameters.
- Work further with the dataset to see if more balanced data for training can be created.

Conclusion

- Using MNIST we were able to reconstruct all classes on partially trained classifiers and from latent extracted form AutoEncoders and VAEs
- Using resnet18 we were able to build models that can make representations of Peruvian ice core images.
- Encountered and overcame problems while training VAEs.
- Fitting less to the artificial set seemed to help the VAE classes.
- Tried clustering and different (V)AE architectures.
- Were able to use representations and dimensionality reduction to identify main classes and outlier images.





Low circularity image surrounded by ones with high circularity. (Outlier)



Local area of small image "light blobs"







QCY_25_7_5_401 Why?





QCY_25_7_5_401.png

QCY_26_4_7_342 Why?



Image Index = 4618

QCY_26_4_7_342.png

QCY_32_3_4_1668

This looks to be a separate cluster and the selected images is an example of it



Image Index = 1551



QCY_23_3_4_1668.png

OCY_26_3_6_931

Visually distinct large object. (15th largest object)





QCY_26_3_6_931.png

QYC_22_1_1_159

Small image (Outlier)



QCY_26_2_5_393

Highest loss for on VAE





QCY_26_2_5_393.png

QCY_26_4_8_1098

Second highest loss on VAE





QCY_26_4_8_1098.png

QCY_25_6_6_262

Previously clustered differently



Appendix

Are you looking for more?

Plain plot



t-Sne for different (V)AEs latent spaces



Refound largest images



CNN classifier - Area 3



CNN classifier - Area 4





Between area 3 and 4



CNN classifier - Area 6



Area 7

Low area





VAE, planar flow Area 1



VAE, planar flow Area 2






Between area 1 and 3



VAE, planar flow Area 4



VAE, planar flow Area 5





VAE, planar flow Area 6



Peruvian Dataset with Labels obtained through the CNN-Classifier

