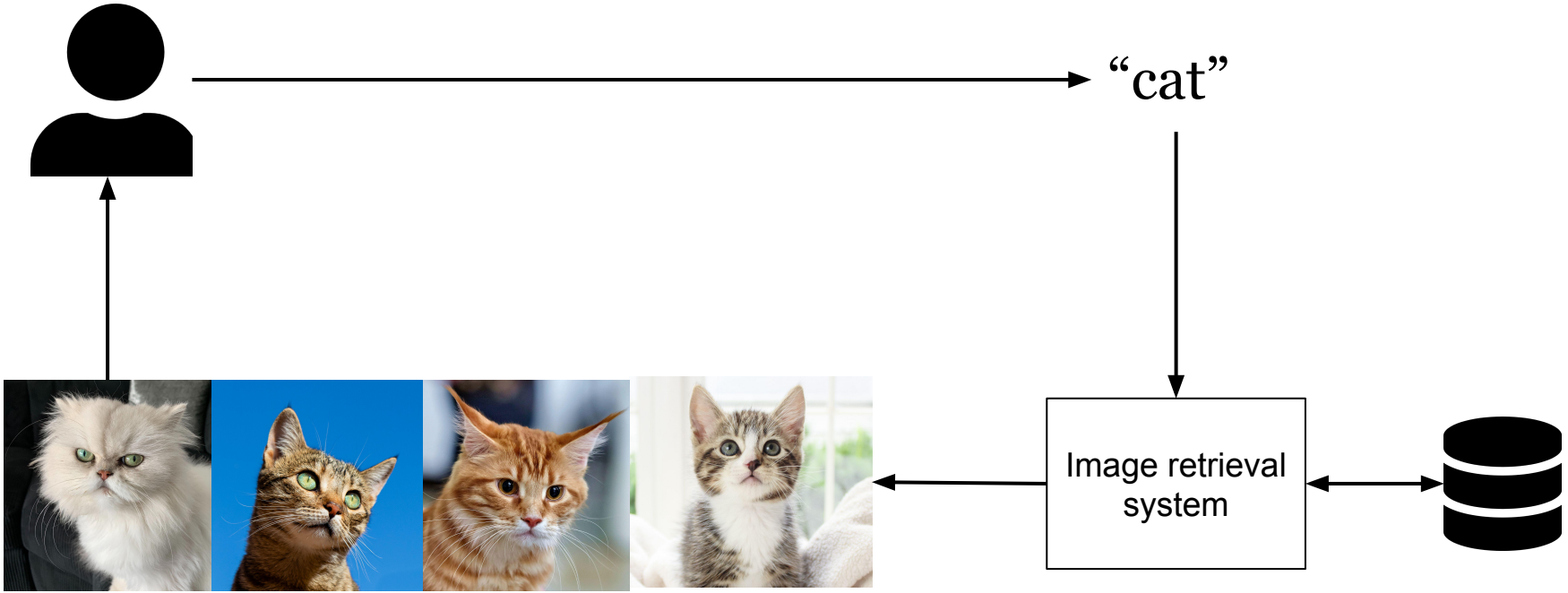
Counterfactual Content-Based Image Retrieval

via Variational Twin Networks and VQ-VAEs

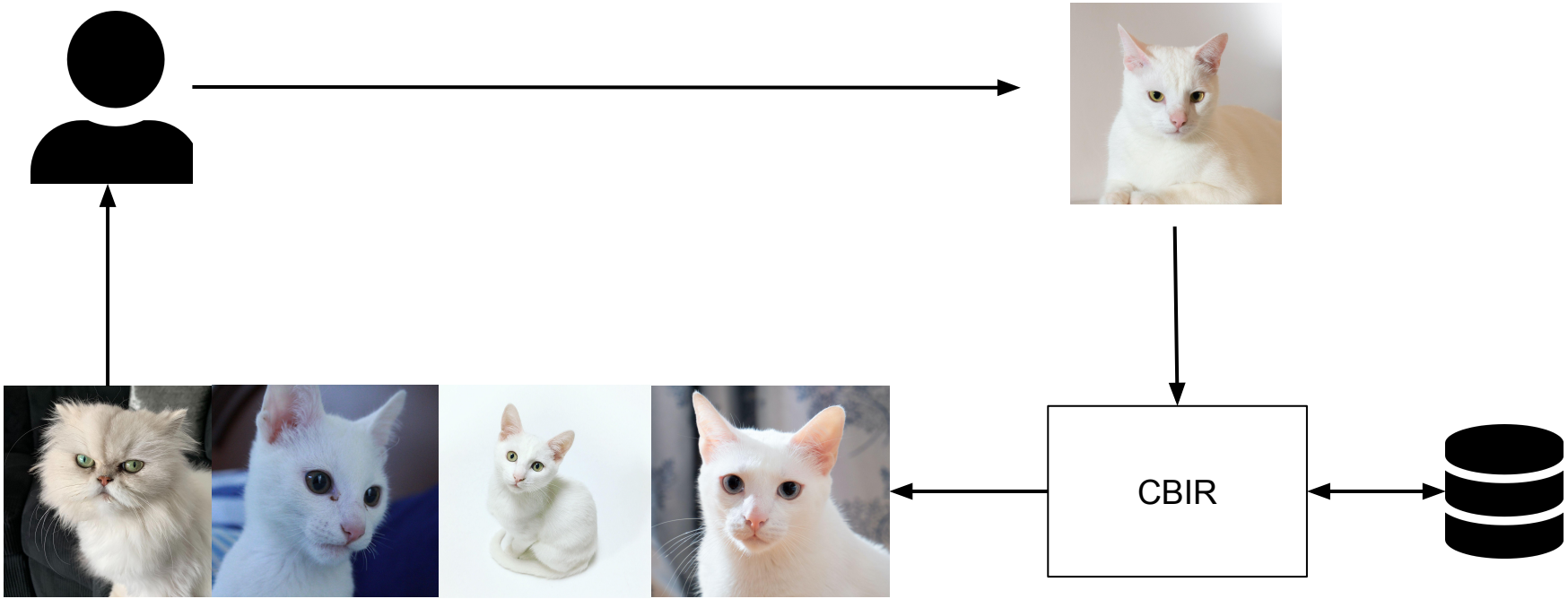
Author: Matteo Bilardi
Supervisors: Athanasios Vlontzos, Bernhard Kainz

- Image retrieval
- Content-based
- Counterfactual

Image retrieval



Content-based image retrieval (aka visual search)

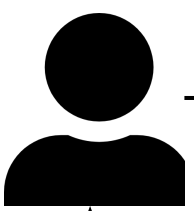


Counterfactual reasoning

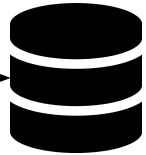
Hypothesise how the world would have looked like today, had something about the past been different, given the hindsight that we now have gained from the present.

Counterfactual CBIR

what if *this* cat had been black?

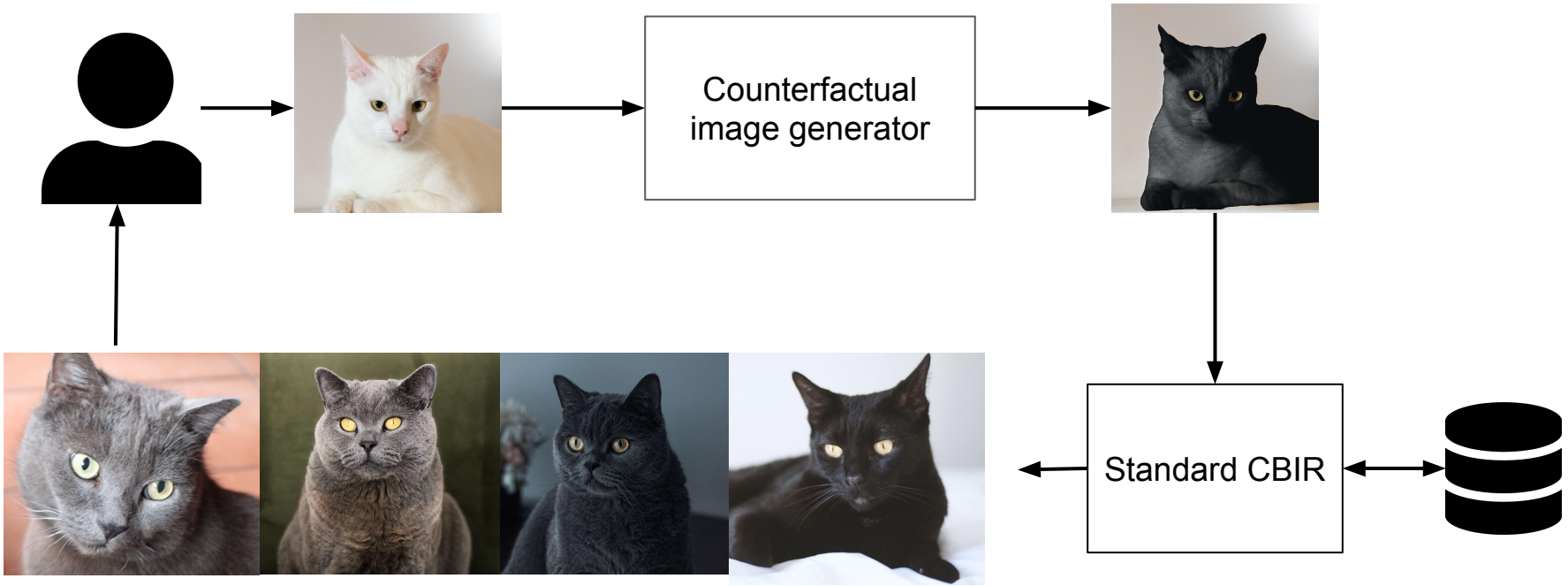


Counterfactual CBIR



Counterfactual CBIR by sample-similarity

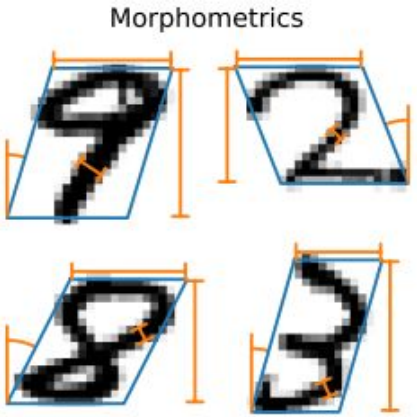
what if *this* cat had been black?



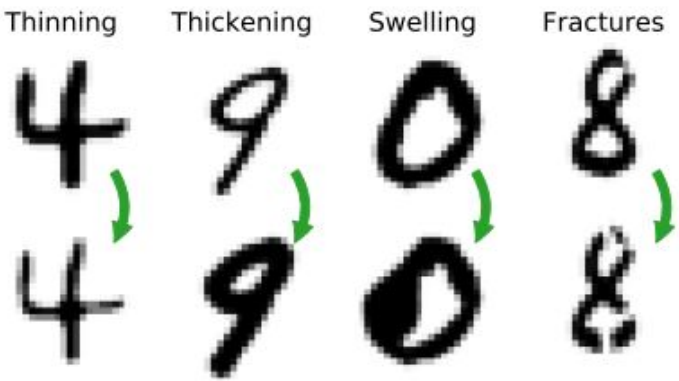
How to generate counterfactual images?

Dataset: Morpho-MNIST

Morphometrics include intensity, thickness, height, width, slant, area



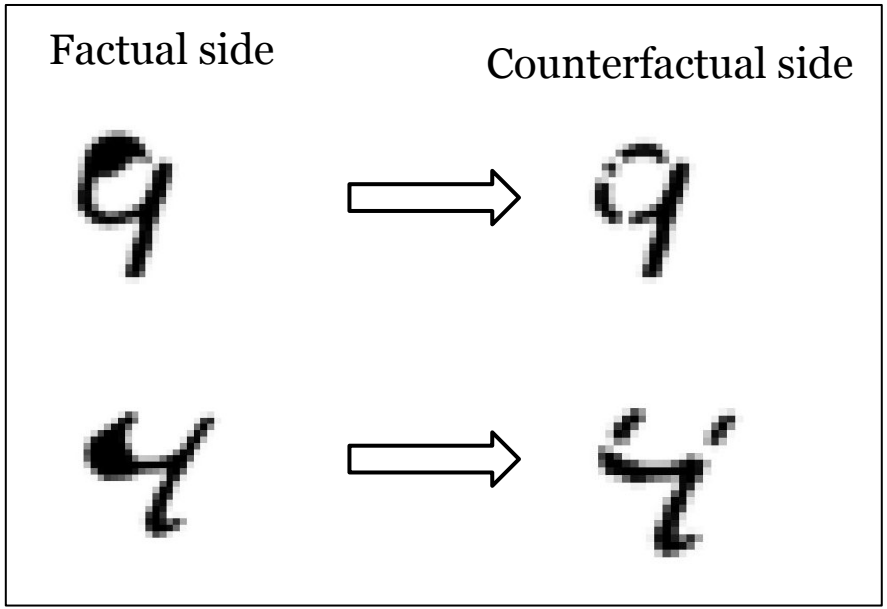
Perturbations



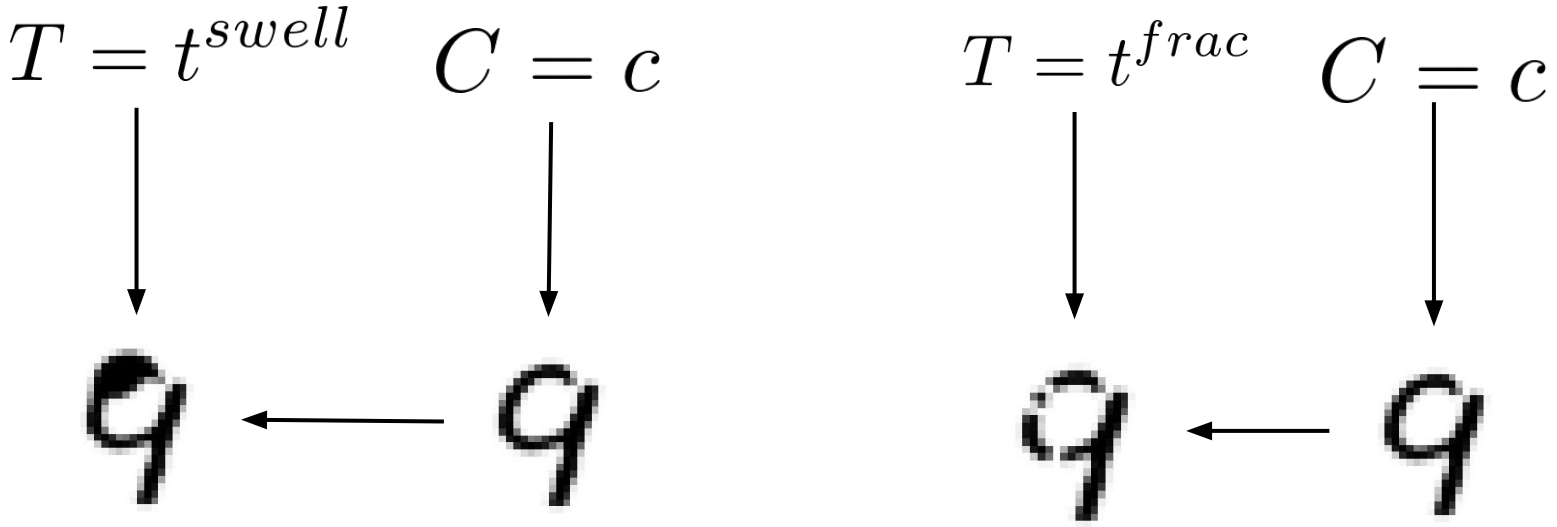
Taken from [3]

[3]: Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative Assessment and Diagnostics for Representation Learning, October 2019. URL http://arxiv.org/abs/1809.10780

Generating counterfactuals: the task

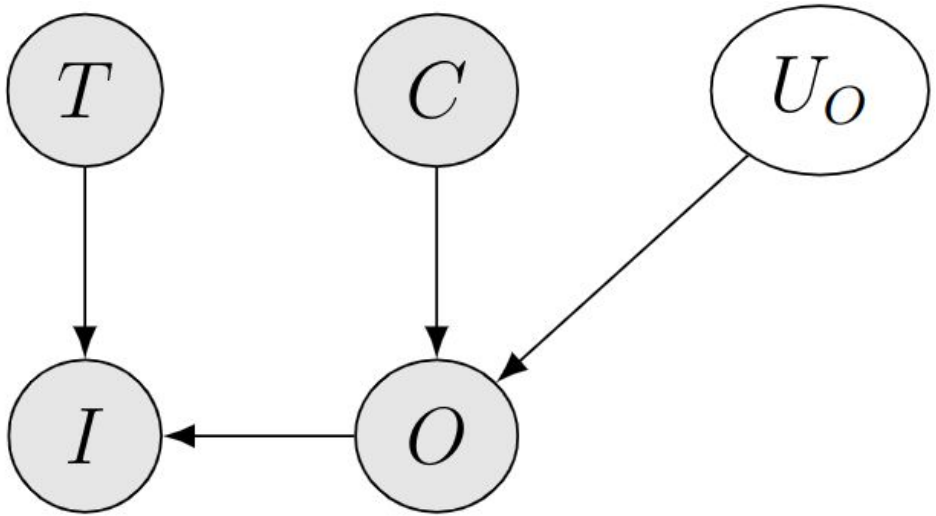


Generating counterfactuals: the background knowledge

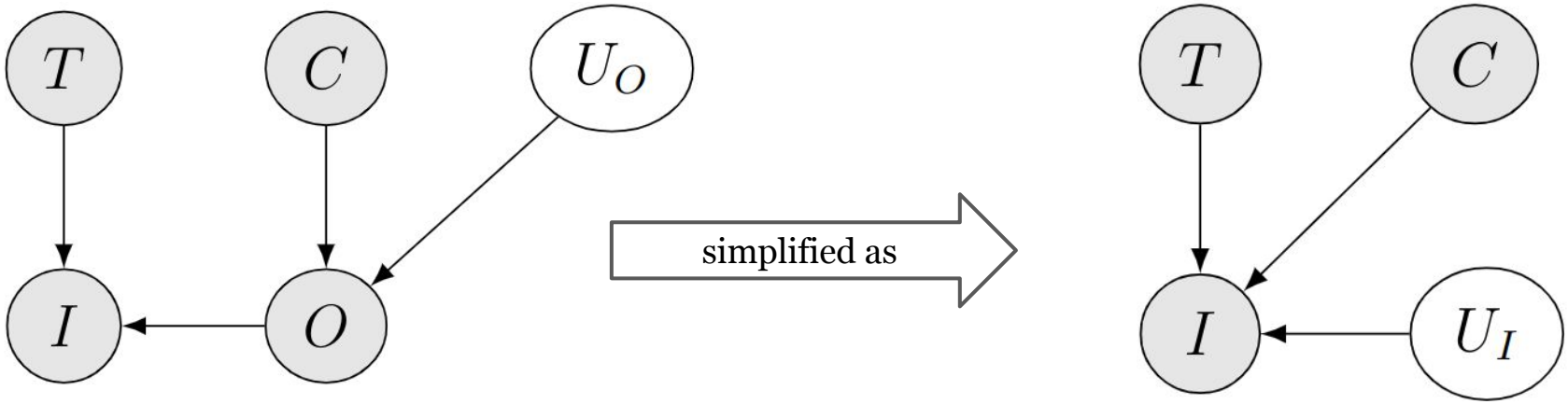


(The covariates *C* include the digit label and morphometrics of the original image)

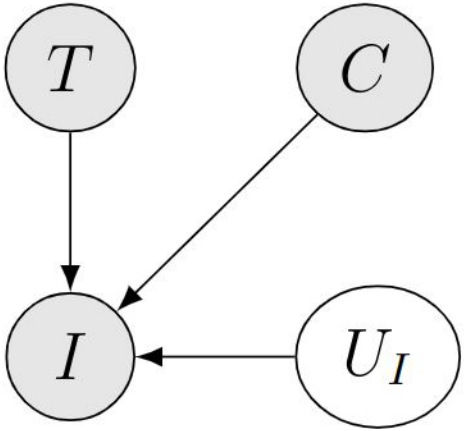
Generating counterfactuals: the (full) causal graph



Generating counterfactuals: the (assumed) causal graph

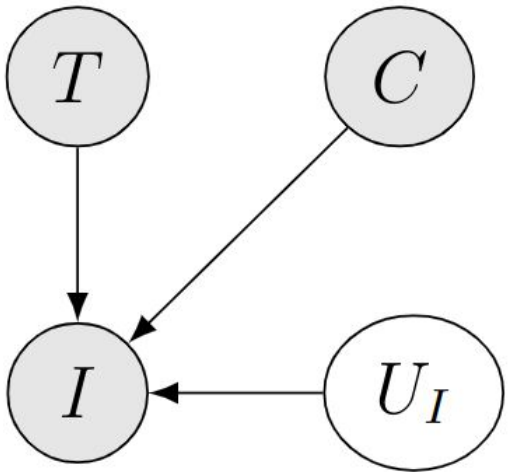


Generating counterfactuals: the (assumed) causal graph



Causal graph alone does not fully describe the generative process.

Generating counterfactuals: (fully-specified) structural causal model



- Exogenous (unobservable) variables $U = \{U_I\}$
- Endogenous (observable) variables $V=\{I, T, C\}$
- Prior over exogenous variables $P(U)$
- Functions to determine the endogenous variables from their causal parents:

$$I := F_I(U_I, T, C)$$

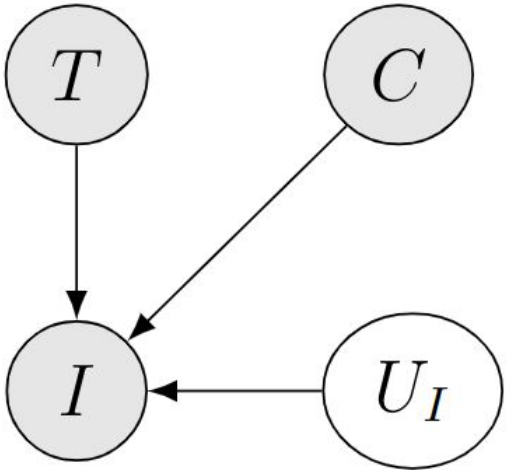
[16] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's Hierarchy and the Foundations of Causal Inference. Technical report, 2021. URL https://causalai.net/r60.pdf

If we had a fully-specified structural causal model, how could we generate counterfactuals?

Sampling counterfactuals via abduction-action-prediction

Recall the user gives $i^{swell}, t^{swell}, t^{frac}, c$

so we know $\exists u_I^{true} : i^{swell} = F_I(u_I^{true}, t^{swell}, c)$



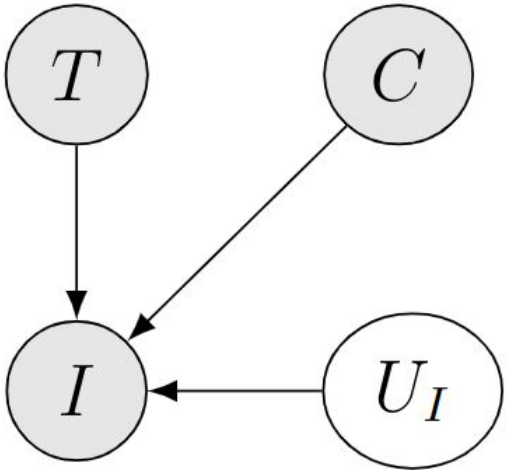
Sampling counterfactuals via abduction-action-prediction

Recall the user gives $i^{swell}, t^{swell}, t^{frac}, c$

so we know $\exists u_I^{true} : i^{swell} = F_I(u_I^{true}, t^{swell}, c)$

$$u_I \sim P(U_I | T = t^{swell}, C = c, I = i^{swell})$$

1. abduction



Sampling counterfactuals via abduction-action-prediction

Recall the user gives $i^{swell}, t^{swell}, t^{frac}, c$

so we know $\exists u_I^{true} : i^{swell} = F_I(u_I^{true}, t^{swell}, c)$

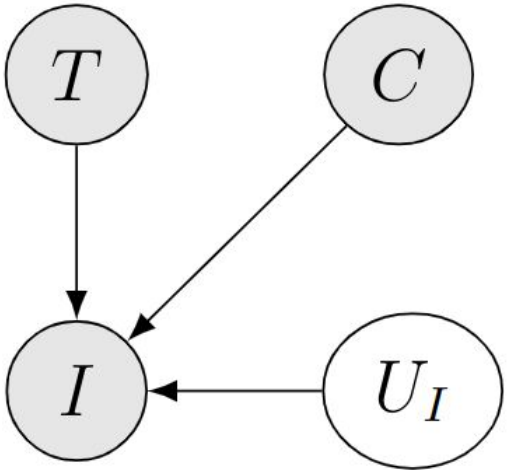
$$u_I \sim P(U_I | T = t^{swell}, C = c, I = i^{swell})$$

$$i^{frac} := F_I(u_i, t^{frac}, c)$$

1. abduction

2. action

3. prediction



Sampling counterfactuals via abduction-action-prediction

Recall the user gives $i^{swell}, t^{swell}, t^{frac}, c$

so we know $\exists u_I^{true} : i^{swell} = F_I(u_I^{true}, t^{swell}, c)$

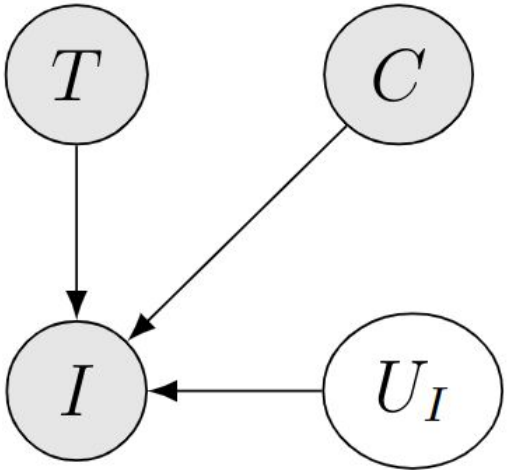
$$u_I \sim P(U_I | T = t^{swell}, C = c, I = i^{swell})$$

$$i^{frac} := F_I(u_i, t^{frac}, c)$$

1. abduction

2. action

3. prediction

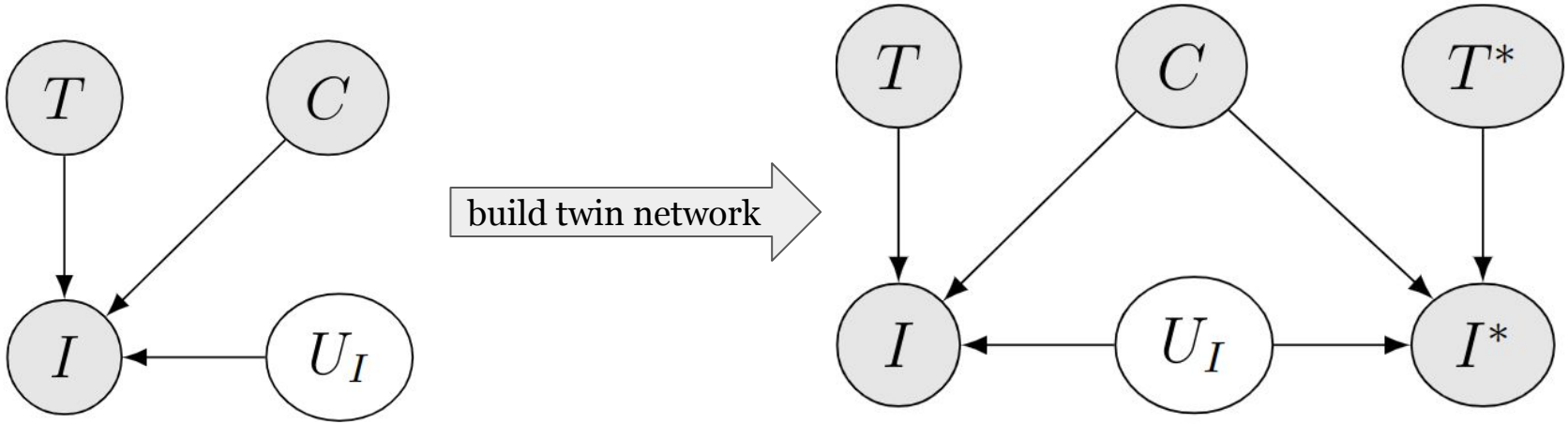


$$i^{frac} \sim P(I_{T:=t^{frac}} | T = t^{swell}, C = c, I = i^{swell})$$

There is an alternative to standard abduction-action-prediction inference, which can save both memory and computational time by avoiding the abduction step [20]: the twin network.

[20] Logan Graham, Ciarán M Lee, and Yura Perov. Copy, paste, infer: A robust analysis of twin networks for counterfactual inference. page 9. URL https://cpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/a/238/files/2019/12/Id_65_final.pdf

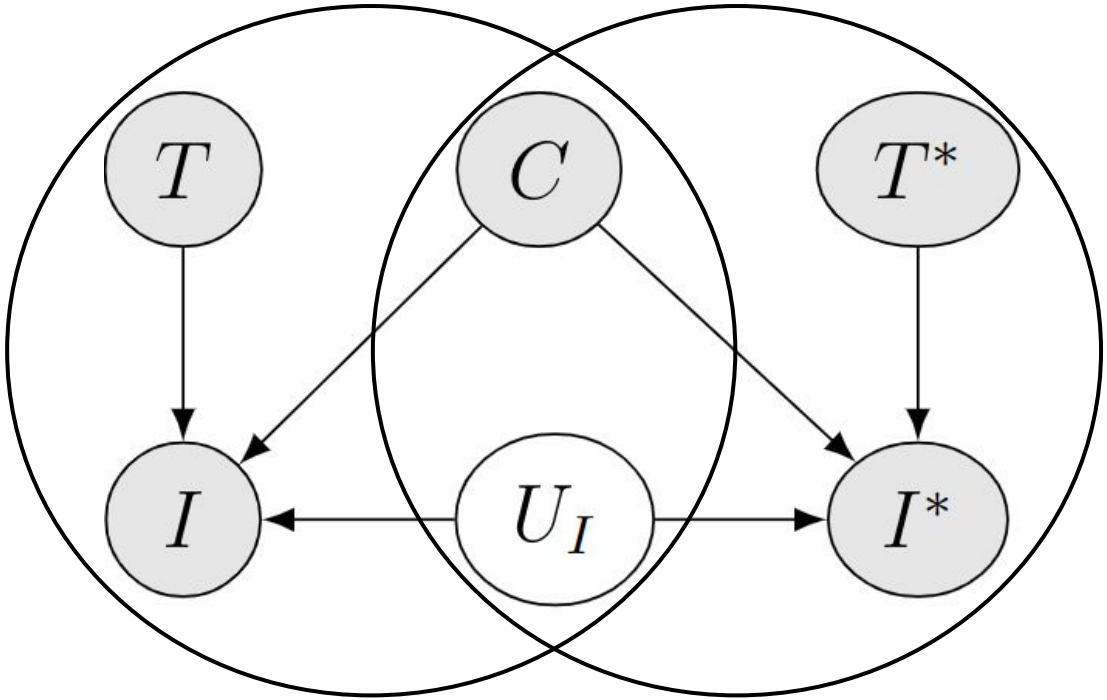
Generating counterfactuals via twin networks: construction



- Duplicate treatments and their causal descendants to add counterfactual nodes.
- Duplicate mechanisms, so $F_I = F_{I^*}$

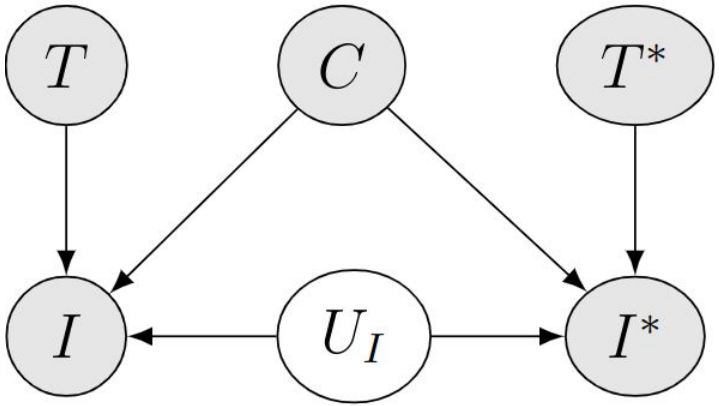
[15] Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, U.K. ; New York, second edition, 2009. ISBN 978-0-521-89560-6.

Generating counterfactuals via twin networks: construction



Factual world $\cup$ Counterfactual world

Generating counterfactuals via twin networks: inference (without abduction)

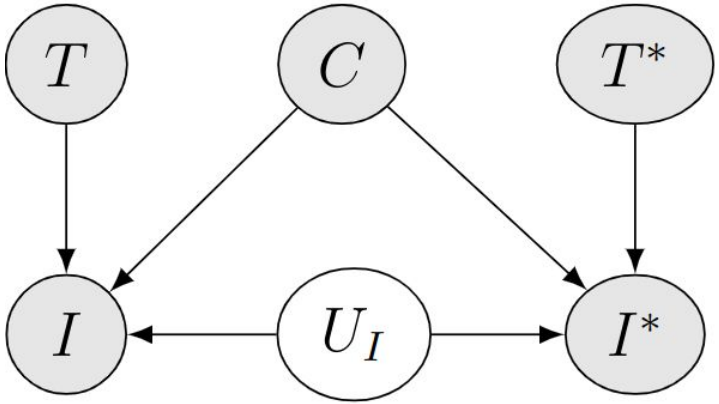


- Given a query from the user $i^{swell}, t^{swell}, t^{frac}, c$
- Sample many $u_I \sim U_I$
- Evaluate $I := F_I(u_I, t^{swell}, c), I^* := F_{I^*}(u_I, t^{frac}, c)$
- if $I = i^{swell}$ then I^* is a valid counterfactual sample for the factual image

[15] Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, U.K. ; New York, second edition, 2009. ISBN 978-0-521-89560-6.

- Full SCM is unknown for most real world settings
- Causal graph may be built from experts' knowledge
- But we still need to learn the functional mechanisms…

Deep twin networks: twin networks with neural mechanisms



$$I := F_I(U_I, T, C) \triangleq NN_{\theta_1}([U_I, T, C])$$
$$I^* := F_{I^*}(U_I, T^*, C) \triangleq NN_{\theta_2}([U_I, T^*, C])$$

- Mechanisms for all endogenous variables are represented as neural networks
- Training requires ground truth counterfactuals (which can be imputed in real world settings)
- Sampling/Inference is identical to a standard twin network

[2] Athanasios Vlontzos, Bernhard Kainz, and Ciaran M. Gilligan-Lee. Estimating the probabilities of causation via deep monotonic twin networks. arXiv:2109.01904 [cs], September 2021. URL http://arxiv.org/abs/2109.01904

Shortcomings of deep twin networks training

- Reynaud et al. [4]: sample noise on each forward pass and multiply with covariates + treatment for diversity

- Vlontzos et al. [2]: sample noise before training and associate one noise value to each dataset item

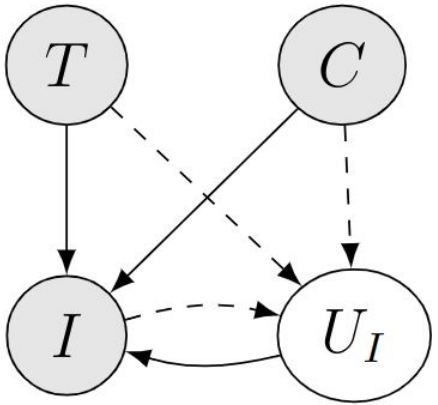
- Both methods fail to assign high probability outcomes to high probability noise under P(U)

[2] Athanasios Vlontzos, Bernhard Kainz, and Ciaran M. Gilligan-Lee. Estimating the probabilities of causation via deep monotonic twin networks. arXiv:2109.01904 [cs], September 2021. URL http://arxiv.org/abs/2109.01904
[4] Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'ARTAGNAN: Counterfactual Video Generation, June 2022. URL http://arxiv.org/abs/2206.01651

Can we train deep twin networks that deal with noise variables in a more principled manner?

DeepSCM: (variational) abduction-action-prediction

- Pawloski et al. [1] propose a method to learn mechanism of a standard SCM as NN components



- Use a variational approach to abduce the noise, e.g. by using a conditional encoder parametrising a variational distribution

$$P(U_I|I,T,C) \approx Q(U_I|I,T,C) \triangleq Q(U_I; e_\psi(I,T,C))$$

- The mechanism for *I* can be a conditional decoder network

$$I := F_I(U_I,T,C) \triangleq d_\theta(U_I,T,C)$$

- Training objective: maximise ELBO as in VAEs

[1]: Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. arXiv:2006.06485 [cs, stat], October 2020. URL http://arxiv.org/abs/2006.06485

This project proposes the application of a variational approach similar to DeepSCM [1] to the architecture of a deep twin networks for twin network inference.

[1]: Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. arXiv:2006.06485 [cs, stat], October 2020. URL http://arxiv.org/abs/2006.06485

Variational twin network (VTN)

Applying the variational approach of DeepSCM to the training of deep twin networks:

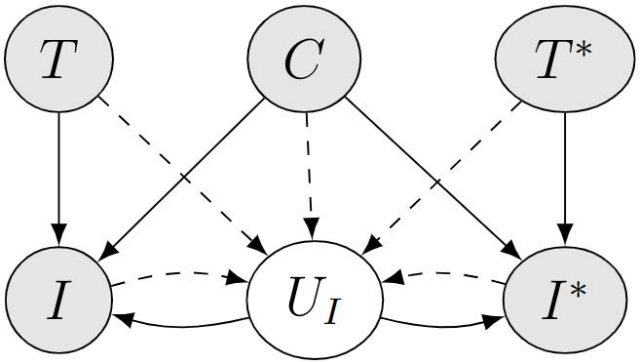
- Use 1 encoder network to parametrise a variational distribution:

$$Q(U|I, I^*, T, T^*, C) \approx Q(U; e_\psi(I, I^*, T, T^*, C))$$

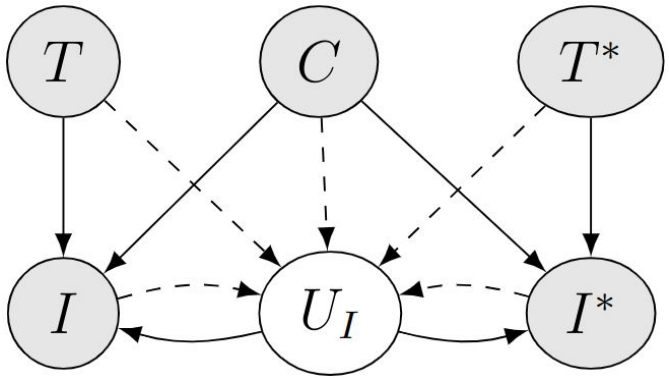
- Use 2 decoder networks for each factual-counterfactual variable pair, equivalent to the NNs in DTNs:

$$I := F_I(U_I, T, C) \triangleq d_\theta(U_I, T, C)$$

$$I^* := F_{I^*}(U_I, T^*, C) \triangleq d_{\theta^*}(U_I, T^*, C)$$



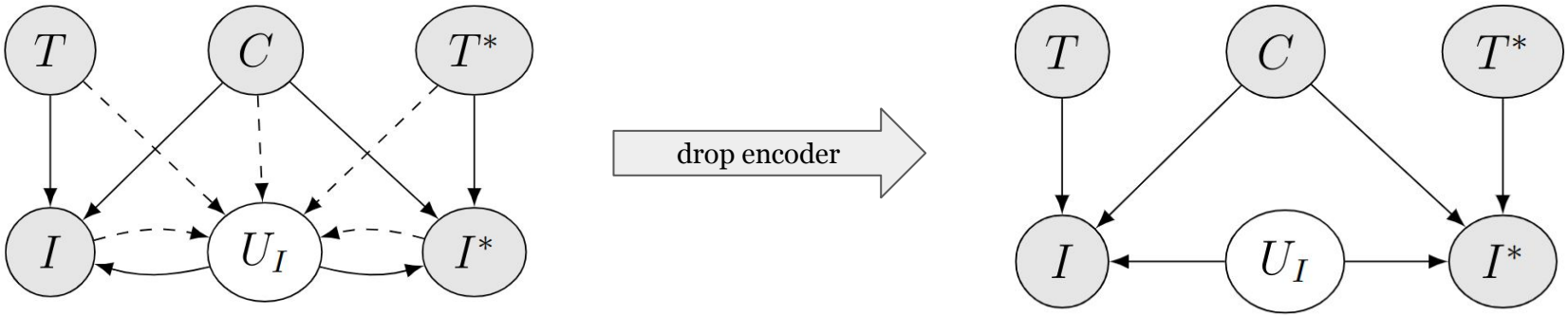
VTN training objective: ELBO



parametrised by decoder networks

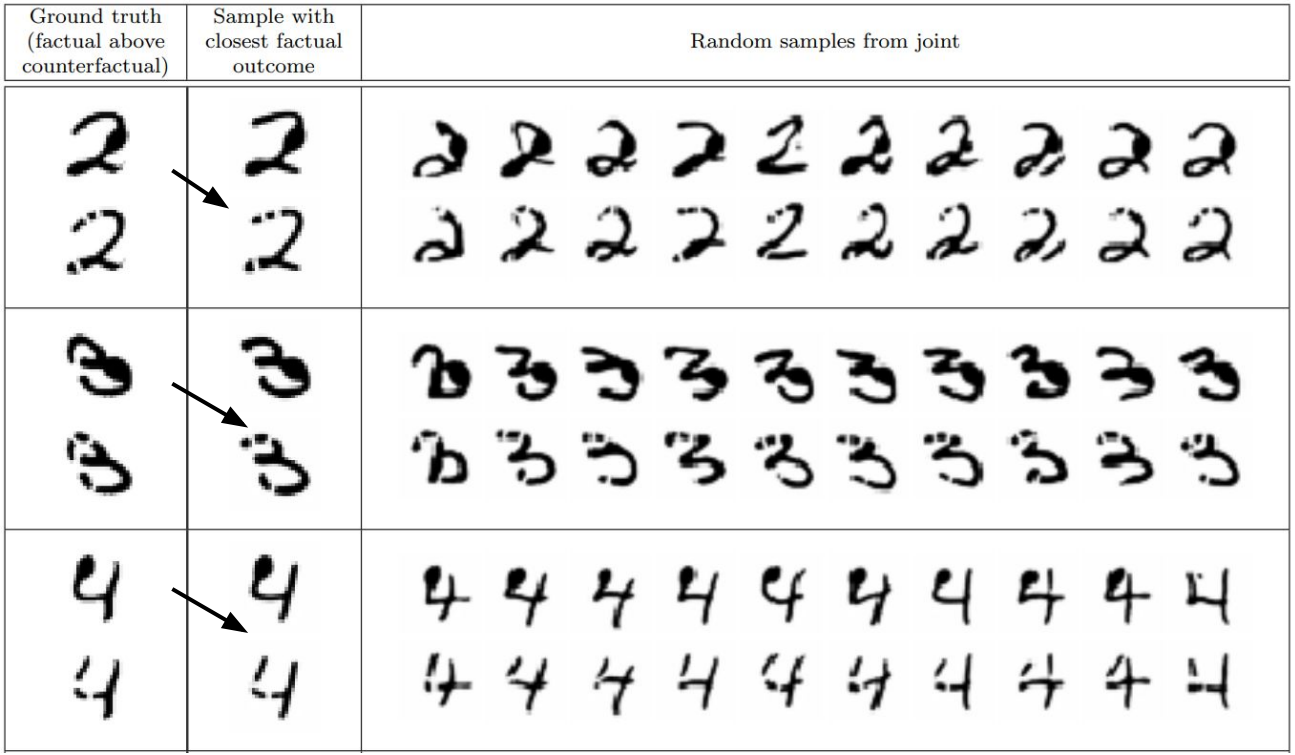
$$ELBO(i, i^*, t, t^*, c)$$
$$\triangleq \mathbb{E}_{Q_\psi(U_I|i,i*,t,t^*,c)}[\log P_\theta(i|U_I, t, c) + \log P_{\theta^*}(i^*|U_I, t^*, c)]$$
$$-KL[Q_\psi(U_I|i, i^*, t, t^*, c)||P(U_I)]$$

VTN: generating samples after training



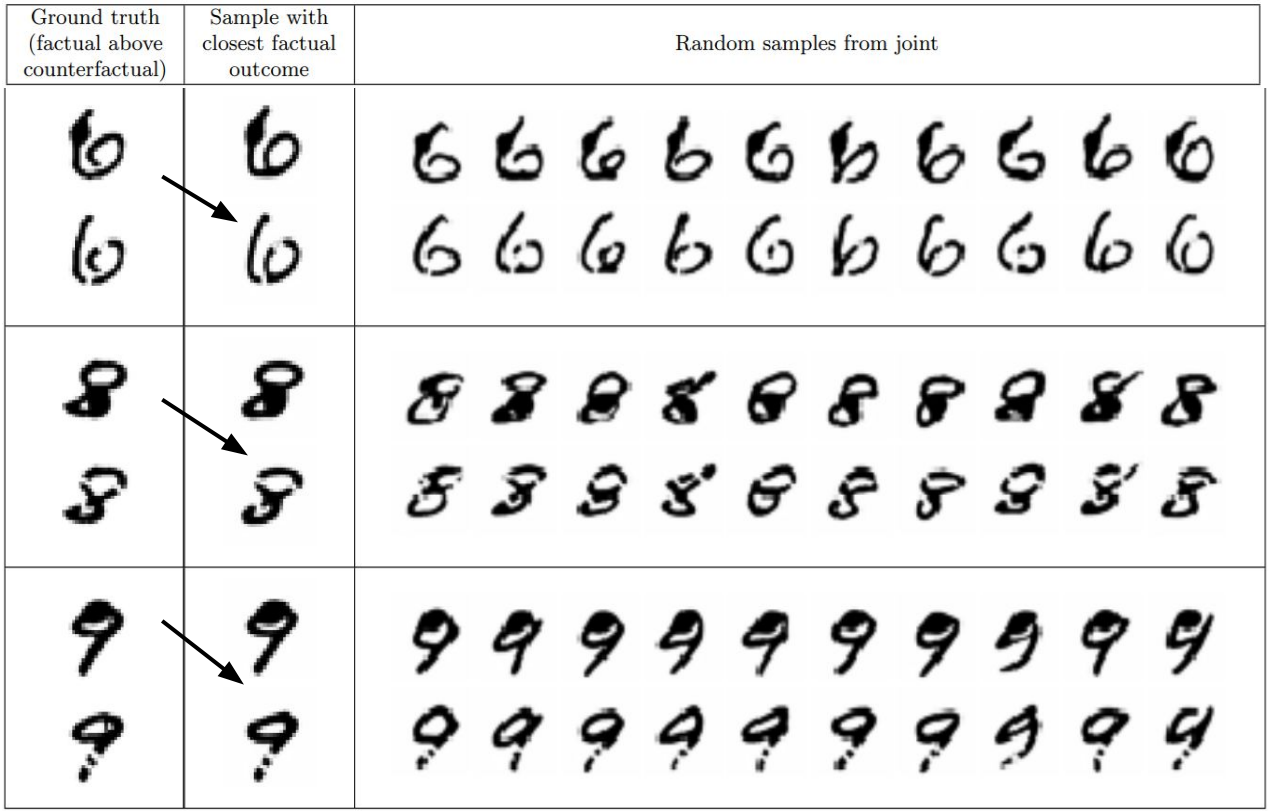
- A DeepSCM uses the encoder network both for training and inference
- The VTN uses the encoder only as an aid in training
- After training, the encoder can be dropped and the VTN becomes a DTN

VTN: qualitative image generation results



- selected counterfactual image is very close to the ground truth counterfactual
- images in each pair seem to come from the same original image
- treatments and covariates are largely respected
- diversity is present (conditional on T and C)

VTN: qualitative image generation results



- selected counterfactual image is very close to the ground truth counterfactual
- images in each pair seem to come from the same original image
- treatments and covariates are largely respected
- diversity is present (conditional on T and C)

Vector Quantised-VAE

- VQ-VAE = autoencoder with vector quantiser bottleneck
- VTN implemented follows Reynolds et al. [4] in using a VQ-VAE to decrease image dimensionality
- In the sample-similarity retrieval case, it can improve quality of distance computations
- discrete vs. continuous latent space

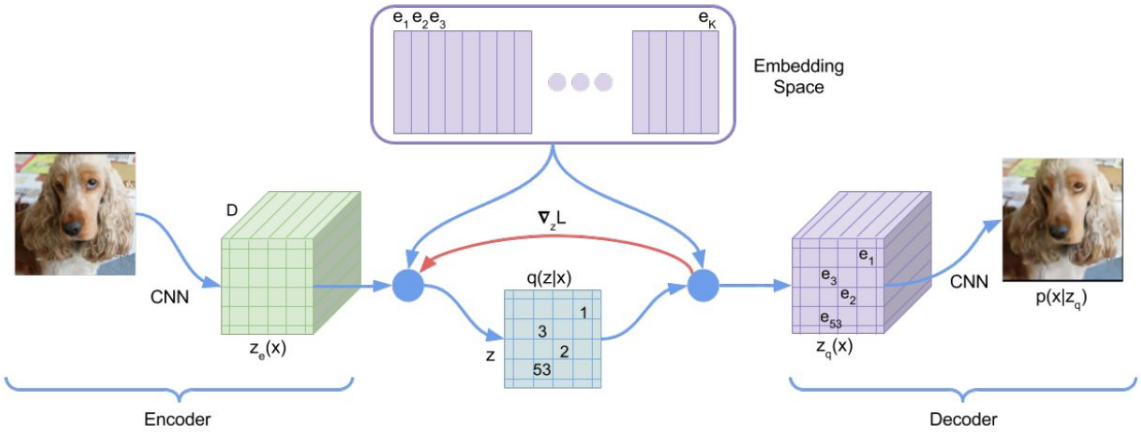


Diagram taken from [35]

[35] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. arXiv:1711.00937 [cs], May 2018. URL http://arxiv.org/abs/1711.00937

[4] Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'ARTAGNAN: Counterfactual Video Generation, June 2022. URL http://arxiv.org/abs/2206.01651

Vector Quantised-VAE

- VQ-VAE = autoencoder with vector quantiser bottleneck
- VTN implemented follows Reynolds et al. [4] in using a VQ-VAE to decrease image dimensionality
- In the sample-similarity retrieval case, it can improve quality of distance computations
- discrete vs. continuous latent space

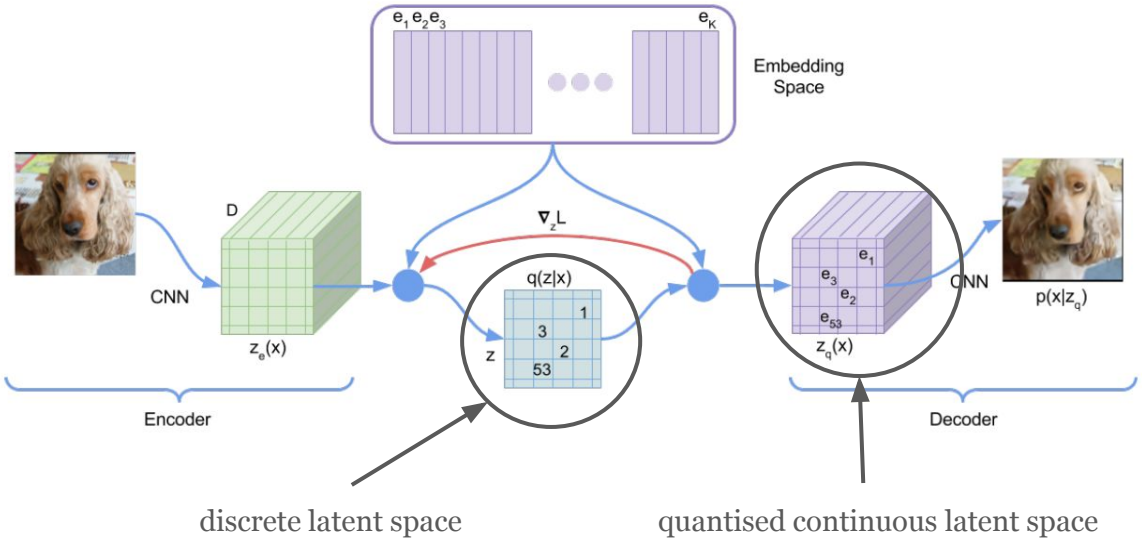


Diagram taken from [35] (except for grey overlays describing latent spaces, which are ours)

[35] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. arXiv:1711.00937 [cs], May 2018. URL http://arxiv.org/abs/1711.00937
[4] Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'ARTAGNAN: Counterfactual Video Generation, June 2022. URL http://arxiv.org/abs/2206.01651

VTN: quantitative image generation results

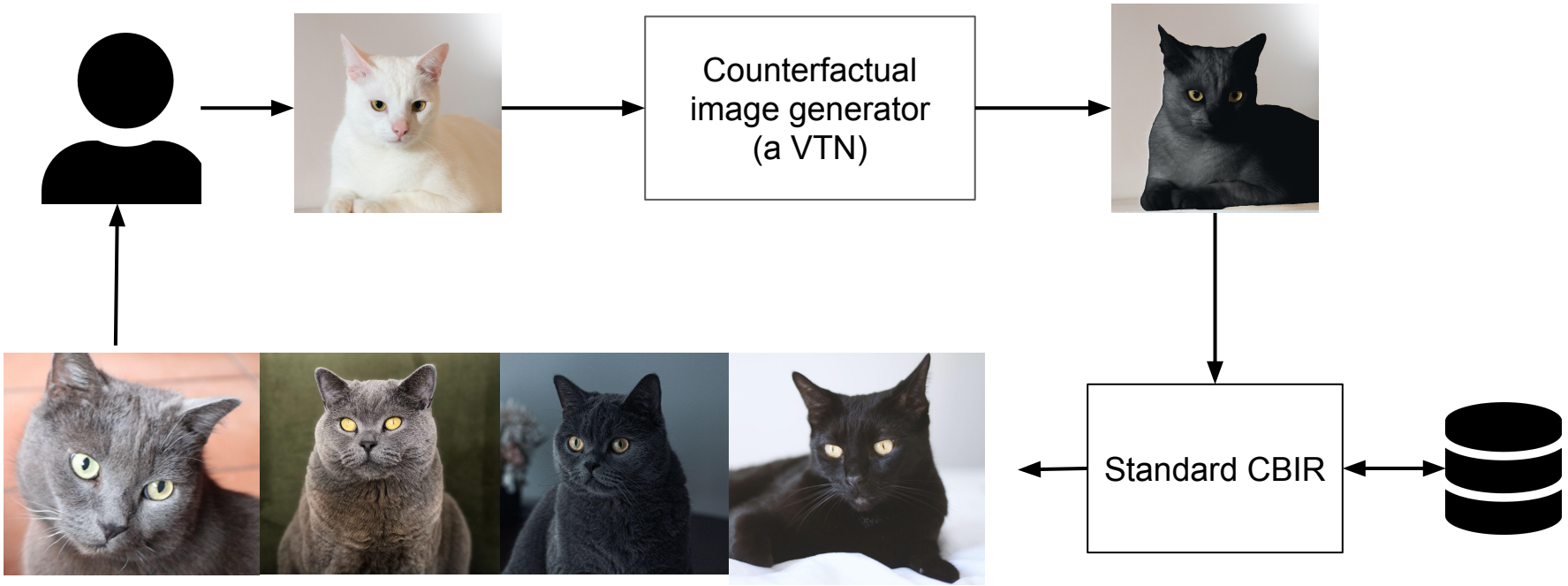
Metric	SSIM(i_{gt}, i_{rec})		SSIM(i_{rec}, i_{pred})		SSIM(i_{gt}, i_{pred})	
Side	factual	counterfact.	factual	counterfact.	factual	counterfact.
DTN [4]	0.9308	0.9308	0.6759	0.6759	0.6707	0.6705
Continuous-VTN	0.9929	0.9923	0.7447	0.7339	0.7438	0.7331
Categorical-VTN	0.9929	0.9923	**0.7987**	**0.7909**	**0.7979**	**0.7900**

- 79% SSIM for best model
- categorical > continuous
- negligible reconstruction loss from VQ-VAE compression

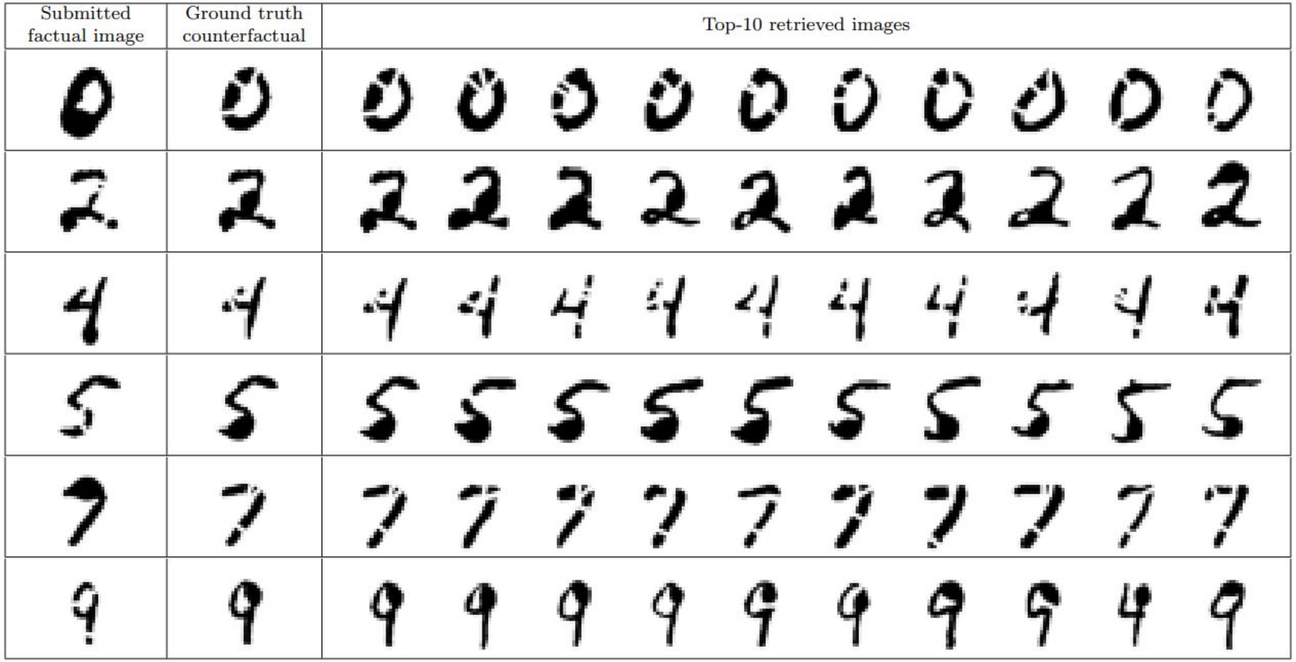
[4]: Results for DTN are from Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'ARTAGNAN: Counterfactual Video Generation, June 2022. URL http://arxiv.org/abs/2206.01651.

Counterfactual CBIR by sample-similarity

what if *this* cat had been black?



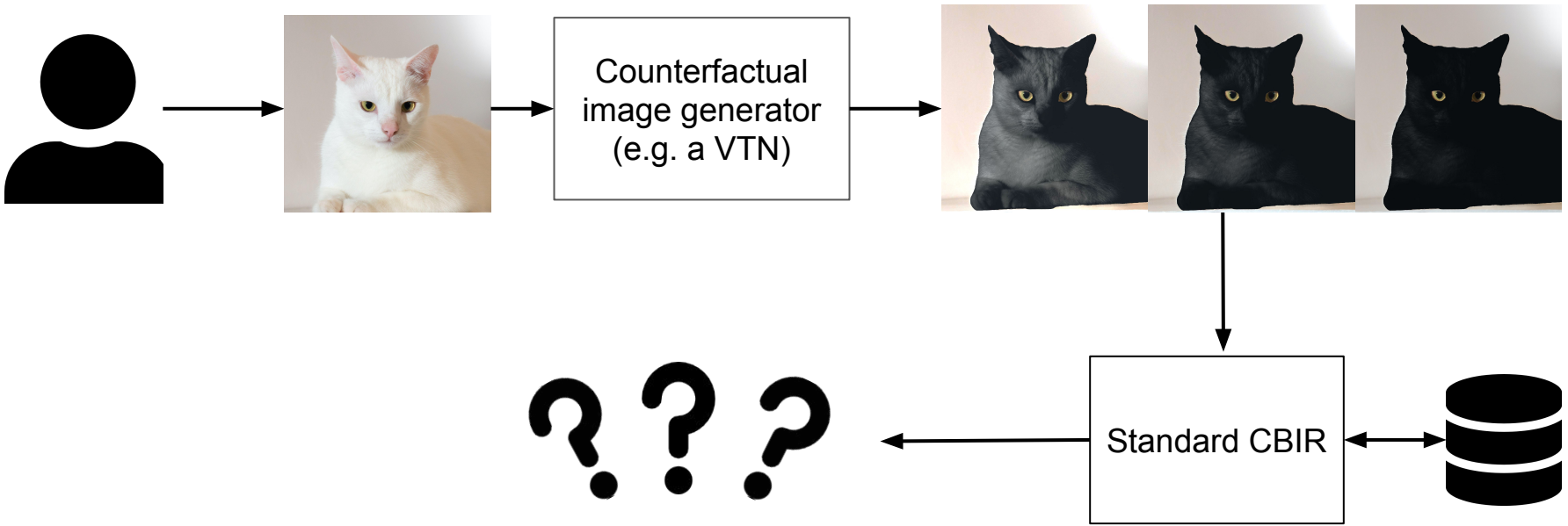
CCBIR: qualitative results on a sample of queries



- top result is ground truth counterfactual
- remaining results are very similar to the ground truth counterfactual

A potential issue with sample-similarity: multiple counterfactuals

what if *this* cat had been black?



Instead of sampling likely counterfactuals and ordering by similarity, can’t we just compute the probability that each image in the database is a counterfactual for the submitted factual image and query?

Ranking by counterfactual probability

Assume the user has submitted a query $i^{swell}, t^{swell}, t^{frac}, c$ and i_n is any image from gallery dataset

$$P_{\mathcal{M}}(I_{T:=t^{frac}} = i_n | T = t^{swell}, C = c, I = i^{swell})$$

relevance score that we wish to rank images by

Ranking by counterfactual probability

Assume the user has submitted a query $i^{swell}, t^{swell}, t^{frac}, c$ and i_n is any image from gallery dataset

$$P_{\mathcal{M}}(I_{T:=t^{frac}} = i_n | T = t^{swell}, C = c, I = i^{swell})$$

$$P_{\mathcal{M}^{TN}_{T^*:=t^{frac}}}(I^* = i_n | T = t^{swell}, C = c, I = i^{swell})$$

relevance score that we wish to rank images by

identical quantities

Ranking by counterfactual probability

Assume the user has submitted a query $i^{swell}, t^{swell}, t^{frac}, c$ and i_n is any image from gallery dataset

$$P_{\mathcal{M}}(I_{T:=t^{frac}} = i_n | T = t^{swell}, C = c, I = i^{swell})$$

$$P_{\mathcal{M}^{TN}_{T^*:=t^{frac}}}(I^* = i_n | T = t^{swell}, C = c, I = i^{swell})$$

$$P_{\mathcal{M}^{TN}_{T^*:=t^{frac}}}(I = i^{swell}, I^* = i_n | T = t^{swell}, C = c)$$

relevance score that we wish to rank images by

identical quantities

identical orderings over I^*

Ranking by counterfactual probability

Assume the user has submitted a query $i^{swell}, t^{swell}, t^{frac}, c$ and i_n is any image from gallery dataset

$$P_{\mathcal{M}}(I_{T:=t^{frac}} = i_n | T = t^{swell}, C = c, I = i^{swell})$$

$$P_{\mathcal{M}^{TN}_{T^*:=t^{frac}}}(I^* = i_n | T = t^{swell}, C = c, I = i^{swell})$$

$$P_{\mathcal{M}^{TN}_{T^*:=t^{frac}}}(I = i^{swell}, I^* = i_n | T = t^{swell}, C = c)$$

relevance score that we wish to rank images by

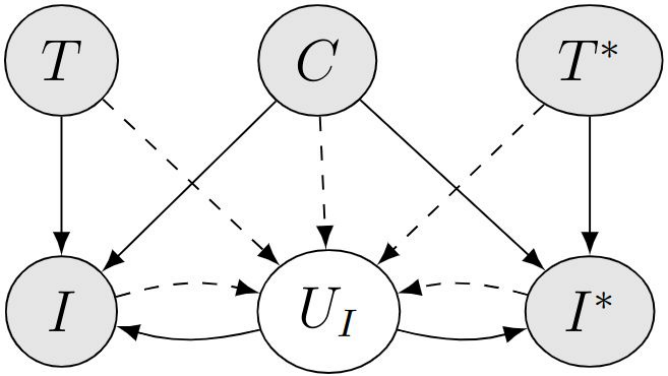
identical quantities

identical orderings over I^*

But all of the above quantities are intractable :(
If only we had a tractable approximation to one of these…

VTN **training objective** *and* **relevance score** for CCBIR: ELBO

Bring back the encoder network!



$$ELBO(i, i^*, t, t^*, c)$$
$$\triangleq \mathbb{E}_{Q_\psi(U_I|i,i*,t,t^*,c)}[\log P_\theta(i|U_I, t, c) + \log P_{\theta^*}(i^*|U_I, t^*, c)]$$
$$-KL[Q_\psi(U_I|i, i^*, t, t^*, c)||P(U_I)]$$

CCBIR: quantitative results

Setup				
Method	CBIR-VQ-VTN		ELBO-VQ-VTN	
Factual	swell	fracture	swell	fracture
Counterfactual	fracture	swell	fracture	swell
Results (%)				
HR@1	98.9	98.3	**99.6**	**99.1**
HR@5	99.3	99.4	**99.8**	**99.8**
HR@10	99.6	99.5	**99.8**	**100.0**
MRR	99.13	98.82	**99.71**	**99.45**
$NDCG_{SSIM}$	**99.87**	**99.89**	99.61	99.71
mAP@All	**26.74**	**25.21**	24.78	23.71
mAP@1000	**75.59**	**69.86**	74.86	68.20

- Both methods are largely successful
- ELBO > sample-similarity for CCBIR
- Performance is generally close
- In standard image retrieval metrics, sample-similarity outperforms ELBO (though irrelevant for CCBIR)

VQ-VAE training techniques evaluated

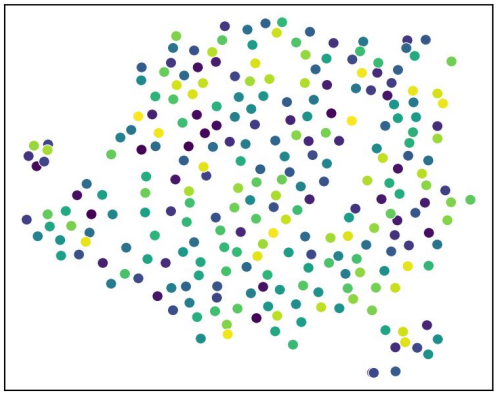
Method	**Potentially desirable effects**
l2-normalisation [82]	cosine distance-implied ordering
product quantisation [13]	larger codebook at small cost
random restarts [12]	codebook collapse mitigation
noisy codeword assignments [11]	continuity between latent space and image space

[82] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized Image Modeling with Improved VQGAN, March 2022. URL http://arxiv.org/abs/2110.04627
[13] Hanwei Wu and Markus Flierl. Learning Product Codebooks Using Vector-Quantized Autoencoders for Image Retrieval. In 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pages 1–5, November 2019. doi: 10.1109/ GlobalSIP45357.2019.8969272
[12] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A Generative Model for Music, April 2020. URL http://arxiv.org/abs/2005.00341
[11]: Harry Coppock and Bjorn Schuller. Vector Quantised-Variational Autoencoders (VQVAEs) for Representation learning. URL https://harrycoppock.com/publication/ 2020-09-01-masterthesis

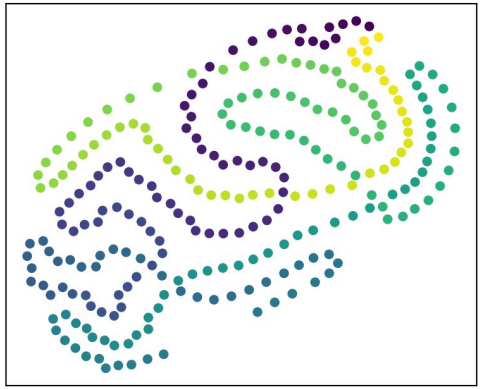
VQ-VAE training techniques: noisy codeword assignments

TSNE visualisations of VQ-VAE codebook:

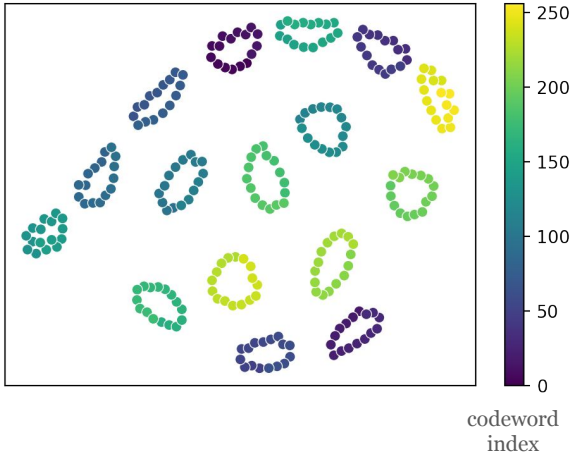
Standard VQ-VAE



Single-ring discrete noise [11]



Multi-ring discrete noise (ours)



[11]: Harry Coppock and Bjorn Schuller. Vector Quantised-Variational Autoencoders (VQVAEs) for Representation learning. URL https://harrycoppock.com/publication/ 2020-09-01-masterthesis

Comparison: other methods for (standard) content-based image retrieval

Method	Results (%)	
	mAP@All	mAP@1000
DeepBit [89]	-	44.53
UTH [90]	-	49.66
PCAH [91]	21.47	63.31
SpeH [92]	24.10	67.60
LSH [93]	31.71	66.23
SphH [94]	34.75	65.45
KMH [95]	35.78	67.62
ITQ [96]	45.37	80.23
DH [83]	46.74	-
DeepQuan [79]	52.54	-
HashGAN [88]	**93.93**	**96.37**
vanilla VQ-VAE	25.95	72.08
PQ-RT-1	49.74	85.10
PQ-RT-8	54.23	80.17

- PQ-RT-8 > 2 * vanilla VQ-VAE (in mAP@All)
- best mAP@All among quantisation methods
- still far below HashGAN's performance

The results for vanilla VQ-VAE, PQ-RT-1 and, PQ-RT-8 have been produced during this project, while the results for other models are taken from table 1 in [88] and table 1 in [79].

[79]: Junjie Chen, William K. Cheung, and Anran Wang. Learning Deep Unsupervised Binary Codes for Image Retrieval. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pages 613–619, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-2-7. doi: 10.24963/ijcai.2018/85. URL https://www.ijcai.org/ proceedings/2018/85.

[88]: Kamran Ghasedi Dizaji, Feng Zheng, Najmeh Sadoughi Nourabadi, Yanhua Yang, Cheng Deng, and Heng Huang. Unsupervised Deep Generative Adversarial Hashing Network. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3664–3673, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi:10.1109/CVPR.2018.00386. URL https://ieeexplore.ieee.org/document/8578484/

Ablation study: results

Model					Results (%)		
ID	PQ/VQ	Distance	Noise rings	Rand. restarts threshold	mAP@All	mAP@1000	$NDCG_{SSIM}$
1	VQ	Euclidean	n/a	n/a	25.95	72.08	98.45
2	VQ	Cosine	n/a	n/a	23.47	66.59	98.67
3	PQ	Euclidean	n/a	n/a	26.57	71.92	98.54
4	PQ	Cosine	n/a	n/a	24.72	71.59	**98.83**
5	PQ	Euclidean	1	n/a	27.75	73.79	98.40
6	PQ	Cosine	1	n/a	27.83	76.51	98.60
7	PQ	Cosine	8	n/a	26.22	74.31	98.48
8	PQ	Euclidean	n/a	0.85	11.05	10.89	94.78
9	PQ	Cosine	n/a	0.85	24.35	70.44	98.77
10	PQ	Euclidean	1	0.85	18.88	28.91	96.42
11	PQ	Cosine	1	0.85	49.74	**85.10**	97.91
12	PQ	Cosine	8	0.85	47.36	80.12	98.01
13	PQ	Cosine	1	0.95	51.57	79.03	97.68
14	PQ	Cosine	8	0.95	**54.23**	80.17	97.71

Ablation study: insights

Method	**Insight**
l2-normalisation	+ training stability
product quantisation	small + in mAP@All
random restarts	small - in mAP@All
noisy codeword assignments	+ in mAP@All
l2-normalisation + random restarts + noisy-assignments	**doubles** mAP@All!
single-ring vs multi-ring?	Depends on stability of training

VQ-VAE training techniques on CCBIR

Setup								
Method	CBIR-VQ-VTN		CBIR-PQ-VTN		ELBO-VQ-VTN		ELBO-PQ-VTN	
Factual	swell	fracture	swell	fracture	swell	fracture	swell	fracture
Counterfactual	fracture	swell	fracture	swell	fracture	swell	fracture	swell
Results (%)								
HR@1	98.9	98.3	97.8	98.3	**99.6**	**99.1**	99.4	99.0
HR@5	99.3	99.4	99.1	99.4	**99.8**	**99.8**	99.6	99.7
HR@10	99.6	99.5	99.2	99.4	**99.8**	**100.0**	99.8	99.8
MRR	99.13	98.82	98.42	98.81	**99.71**	**99.45**	99.50	99.29
$NDCG_{SSIM}$	**99.87**	**99.89**	99.86	99.89	99.61	99.71	98.44	98.97
mAP@All	26.74	25.21	**27.70**	**25.78**	24.78	23.71	17.91	19.64
mAP@1000	**75.59**	**69.86**	75.05	68.39	74.86	68.20	50.36	52.14

- Improvements in standard CBIR did not translate to the CCBIR setting.
- Hypothesised causes: difficulty in fitting product quantiser, different dimensionality constraints.

Recap of main contributions

- Applied variational approach of DeepSCM to deep twin network to propose the variational twin network model.

- Proposed 2 methods to perform CCBIR: sample-similarity rank & ELBO rank

- Identified a combination of training techniques for VQ-VAEs which more than doubles retrieval performance on a standard image retrieval task

Thank you for listening.
Questions?