Variational Learning with Disentanglement-PyTorch

Amir H. Abdi amirabdi@ece.ubc.ca
Purang Abolmaesumi purang@ece.ubc.ca
Sidney Fels ssfels@ece.ubc.ca

Electrical and Computer Engineering Department, University of British Columbia, Vancouver, Canada

Abstract

Unsupervised learning of disentangled representations is an open problem in machine learning. The Disentanglement-PyTorch library is developed to facilitate research, implementation, and testing of new variational algorithms. In this modular library, neural architectures, dimensionality of the latent space, and the training algorithms are fully decoupled, allowing for independent and consistent experiments across variational methods. The library handles the training scheduling, logging, and visualizations of reconstructions and latent space traversals. It also evaluates the encodings based on various disentanglement metrics. The library, so far, includes implementations of the following unsupervised algorithms VAE, β-VAE, Factor-VAE, DIP-I-VAE, DIP-II-VAE, Info-VAE, and β-TCVAE, as well as conditional approaches such as CVAE and IFCVAE. The library is compatible with the Disentanglement Challenge of NeurIPS 2019, hosted on AICrowd and was used to compete in the first and second stages of the challenge, where it was ranked among the best few participants.

Keywords: Disentanglement, PyTorch, Representation Learning, Total Correlation, Factorization

1. Introduction

There are two overlapping avenues in representation learning. One focuses on learning task-specific transformations often optimized towards specific domains and applications. The other approach learns the intrinsic factors of variation, in a disentangled and task-invariant fashion. The unsupervised disentanglement of latent factors, where changes in a single factor of variation shifts the latent encoding in a single direction, is an open problem of representation learning (Bengio et al., 2013; Lake et al., 2016). Disentangled representations are valuable in few-shot learning, reinforcement learning, transfer learning, as well as semi-supervised learning (Bengio et al., 2013; Peters et al., 2017; Locatello et al., 2019).

In this work, we developed a library based on the functionalities of the PyTorch framework, which facilitates research, implementation, and testing of new variational algorithms focusing on representation learning and disentanglement. The library branches from the Disentanglement Challenge of NeurIPS 2019, hosted on AICrowd (aicrowd.com), and was used to compete in the first and second stages of the challenge where it was highly ranked. The Disentanglement-PyTorch library is released under the GNU General Public License at `https://github.com/amir-abdi/disentanglement-pytorch`.

© 2019 A.H. Abdi, P. Abolmaesumi & S. Fels.

2. Library Features

2.1. Supported Algorithms and Objective Functions

Unsupervised Objectives Currently, the library includes implementations of the following unsupervised variational algorithms: VAE (Kingma and Welling, 2014), β-VAE (Fertig et al., 2018), β-TCVAE (Chen et al., 2018), Factor-VAE (Kim and Mnih, 2018), Info-VAE (Zhao et al., 2017), DIP-I-VAE, and DIP-II-VAE (Kumar et al., 2018). Algorithms are implemented as plug-ins to the variational Bayesian formulation, and are specified by the `loss_terms` flag. As a result, if the loss terms of two learning algorithms (*e.g.*, A and B) were found to be compatible, they can both be included in the objective function with the flag set as [`--loss_terms A B`]. This enables researchers to mix and match loss terms which optimize towards correlated goals.

Conditional and Attribute-variant Objectives The library supports conditional approaches such as CVAE (Sohn et al., 2015), where extra known attributes (*i.e*, labels) are included in the encoding and decoding processes. It also supports IFCVAE, inspired by the IFcVAE-GAN (Creswell et al., 2017), which enforces certain latent factors to encode known attributes using a set of positive (auxiliary) and negative (adversarial) discriminators in a supervised fashion. Thanks to the modular implementation of the library, any of the above-mentioned unsupervised loss terms can be used with conditional and information factoriation approaches to encourage disentanglement across attribute-invariant latents.

2.2. Neural Architectures

Neural architectures and the dimensionality of the data and the latent spaces are configurable and decoupled from the training algorithm. Consequently, new architectures for the encoder and decoder networks, such as the auto-regressive models, and support for other data domains, can be independently investigated.

2.3. Evaluation of Disentanglement

We rely on Google's implementation of the disentanglement metrics to evaluate the quality of the learned representations (Locatello et al., 2019). Thanks to the disentanglement-lib[1] library, the following metrics are currently supported: BetaVAE (Higgins et al., 2017), FactorVAE (Kim and Mnih, 2018), Mutual Information Gap (MIG) (Chen et al., 2018), Interventional Robustness Score (IRS) (Suter et al., 2019), Disentanglement Completeness and Informativeness (DCP) (Eastwood and Williams, 2018), and Separated Attribute Predictability (SAP) (Kumar et al., 2018).

2.4. Miscellaneous Features

Controlled Capacity Increase It is shown that gradually relaxing the information bottleneck during training improves the disentanglement without penalizing the reconstruction accuracy. Following the formulation of Burgess et al. (2018), the capacity, defined as the distance between the prior and the latent posterior distributions and denoted with the variable C, is gradually increased during training.

1. https://github.com/google-research/disentanglement_lib

Reconstruction Weight Scheduler To avoid convergence points with high reconstruction loss, training can be started with more emphasis on the reconstruction and gradually relaxing for the disentanglement term to become more relative.

Dynamic Learning Rate Scheduling All forms of learning rate schedulers are supported. Researchers are encouraged to leverage the dynamic LR scheduling to gradually decrease the rate when the average objective function over the epoch stops its decremental trend.

Logging and Visualization The library leverages the Weights & Biases (W&B)[2] tool to record and visualize the training process and experiments' results. Besides the scalar values, we visualize the attribute and condition traversals, latent factor traversals, and input reconstructions, both as static images (logged via W&B) as well as animated GIFs.

3. Experiments and Results

The β-TCVAE algorithm achieved the best disentanglement results on the `mpi3d_real` dataset in the second stage of the disentanglement challenge. Given the limited 8-hour training time for the challenge, the model was pre-trained on the `mpi3d_toy` dataset (Gondal et al., 2019). The model was trained with the Adam optimizer for 90k iterations on batches of size 64. The β value of the β-TCVAE objective function was set to 2. The learning rate was initialized at 0.001 and reduced on the plateau of the objective function with a factor of 0.95. The capacity parameter, C, was gradually increased from 0 to 25. The dimensionality of the z space was generously set to 20.

The encoder consisted of 5 convolutional layers with strides of 2, kernel sizes of 3×3, and number of kernels gradually increasing from 32 to 256. The encoder ended with a dense linear layer which estimated the posterior latent distribution as a parametric Gaussian. The decoder network consisted of one convolutional followed with 6 deconvolutional (transposed convolutional) layers, with kernel sizes of 4, strides of 2, and the number of kernels gradually decreasing from 256 down to the number of channels of the image space. ReLU activations were used except for the last layers of the encoder and decoder networks.

Model's performance on the unseen objects of `mpi3d_realistic` and `mpi3d_real` datasets are presented in Table 1. The configurations of the two experiments are the same and the model consistently performed better on the `mpi3d_real` dataset. This was unexpected as the model was initialized on the `mpi3d_toy` dataset. Disentanglement performance, on the available samples of the `mpi3d_realistic` dataset, is visualized in Appendix A.

Table 1: Results of the best configuration on DCI, FactorVAE, SAP, MIG, and IRS metrics.

Method	Dataset	**DCI**	**FactorVAE**	**SAP**	**MIG**	**IRS**
β-TCVAE	`mpi3d_realistic`	0.3989	0.3614	0.1443	0.2067	0.6315
β-TCVAE	`mpi3d_real`	0.4044	0.5226	0.1592	0.2367	0.6423

2. https://www.wandb.com

References

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, August 2013.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Antonia Creswell, Yumnah Mohamied, Biswa Sengupta, and Anil A Bharath. Adversarial information factorization, 2017.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Emily Fertig, Aryan Arbabi, and Alexander A. Alemi. β-VAEs can retain label information even at high compression, 2018.

Muhammad Waleed Gondal, Manuel Wüthrich, ore Miladinović, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *arXiv preprint arXiv:1906.03292*, 2019.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, November 2016.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. ISBN 978-0-262-03731-0.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3483–3491. Curran Associates, Inc., 2015.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6056–6065, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information maximizing variational autoencoders, 2017.

Appendix A. Latent Factor Traversal

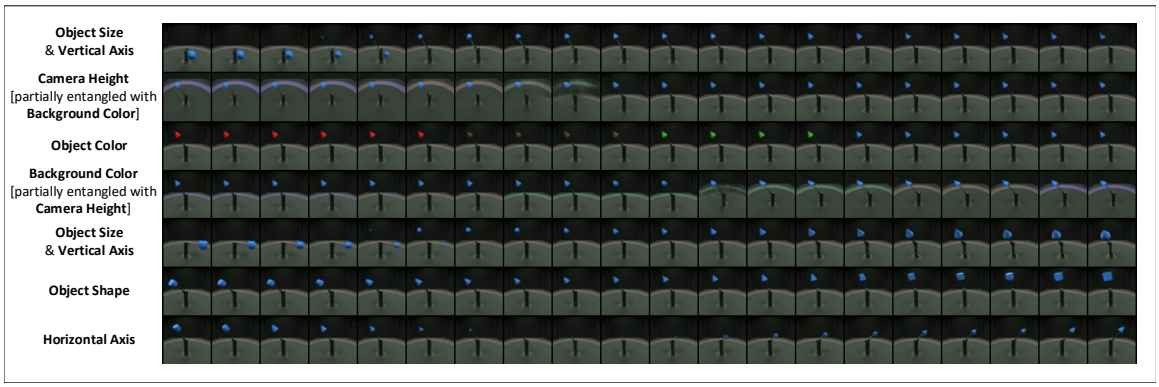

Figure 1: Latent factor traversal of the trained β-**TCVAE** model on a random sample of the `mpi3d_realistic` dataset. As demonstrated, the disentanglement is not complete and some features are encoded in the same latent factor. A latent space of size 20 was used, however, changes in the other 13 latent factors had no effect on the reconstruction; thus, these feature-invariant factors were not included for brevity.