Neural Disentanglement using Mixture Latent Space with Continuous and Discrete Variables

Sourabh Balgi SOURABHBALGI@IISC.AC.IN
and
Ambedkar Dukkipati AMBEDKAR@IISC.AC.IN
Department of Computer Science and Automation
Indian Institute of Science
Bengaluru 560012

Abstract

Recent advances in deep learning techniques have shown extordinary capability of deep neural networks in extracting features required to perform the task at hand. However, the features learnt are relevant only for the initial task at hand. This is due to the fact that the features learnt are usually task specific and do not capture the most general and task agnostic features of the input. On the other hand, humans are excellent task agnositc learners by automatically disentangling the features. Recently variational autoencoders (VAEs) have shown to be the de-facto models to capture the latent variables in a generative sense. As these latent features can be represented as continuous and/or discrete variables, this motivated us to use VAEs with a mixture of continuous and discrete variables for the latent space. We achieve this by performing our experiments using a modified version of *JointVAE* to learn the disentangled features.

Keywords: deep learning, neural disentanglement, unsupervised learning representation, variational auto-encoder

1. Introduction

Feature learning is one of the most fundamental task in machine learning and recently deep learning has made revolutionary advanced in this. What ever the machine learning task at hand, deep neural networks are excellent models for feature extraction from a raw data. But the features extracted or learned are very task specific as one use particular loss functions that are suited for task at hand. For example, cross entropy loss used for multiclass classification problems.

This way of learning performs well only for the particular trained task leading to what is called as a narrow or weak artificial intelligence. However, to achieve the ultimate goal of true or general artificial intelligence, one needs to learn representations in a task agnostic manner. These task agnostic features should be enough to capture all the required information of the given entity.

One such effort made in recent times is towards learning disentangled representations. As Bengio et al. (2013) defines, disentangled representations are the representations where a change in a single unit of the representation corresponds to a change in a single factor of the

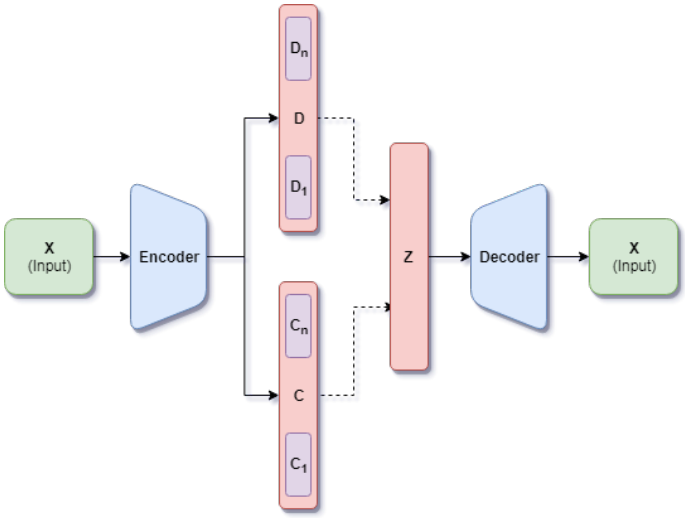

Figure 1: Architecture of JointVAE. C=$[C_1, \ldots, C_{n_c}]$ represents n_c continuous variables and D=$[D_1, \ldots, D_{n_d}]$ represents n_d discrete variables. Z=$[C, D]$ represents final disentangled feature after concatenating continuous and discrete variables.

data while simultaneously being invariant to other factors of variations in the representation. A more formal definition is given in Higgins et al. (2018) as:

> A vector representation is called a **disentangled representation** with respect to a particular decomposition of a symmetry group into subgroups, if it decomposes into independent subspaces, where each subspace is affected by the action of a single subgroup,and the actions of all other subgroups leave the subspace unaffected.

These disentangled representations are very helpful in several downstream machine learning tasks such as transfer learning, multi-task learning and zero-shot learning. These applications of disentangled representations indicate the need for development of new approaches. Since these representations essentially define data generative factors, recently variational auto-encoders (VAEs) Kingma and Welling (2014) have been explored to capture disentangled features in Higgins et al. (2017); Chen et al. (2018); Dupont (2018).

In this work, we experiment with JointVAE Dupont (2018) to explore the disentangled representation for the given dataset Gondal (2019). In the next sections, we discuss our experimental setup and results.

2. A brief description of JointVAE

There are several state-of-the-art variants of VAEs have been reported to extract disentangled representations Higgins et al. (2017); Chen et al. (2018); Dupont (2018). A common assumption in these models is that latent variables follow Gaussian distribution. The reason for this is, (a) this is the most common and simplest form of distribution observed

over multiple naturally occurring datasets, and (b) the assumption of Gaussian distribution helps in simplifying the sampling of latent variables using reparametrization trick.

This assumption of Gaussian distribution holds only when the data follows linear interpolation in both the input feature space and latent variable space. In the case of discrete variables, where we cannot observe the linear interpolation, it is necessary to directly represent the latent feature as a discrete multinominal variables. This is achieved in JointVAE Dupont (2018) by using a mixture of continuous and discrete latent variables to represent the disentangled features. In this model, the continuous variables are assumed to be of Gaussian distribution and the discrete variables are assumed to be multinominal distribution. For continuous latent variables, one can use the normal reparameterization trick for sampling from latent variable. However, in the case of discrete multinominal variable, the sampling is done using Gumbel Softmax trick Jang et al. (2017). This allows us to represent the disentangled features in terms of both continuous and discrete variables without any assumptions, thus utilizing the best of both worlds.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2615–2625, 2018. URL http://papers.nips.cc/paper/7527-isolating-sources-of-disentanglement-in-variational-autoencoders.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*, pages 710–720, 2018.

Wthrich M. Miladinovi . Locatello F. Breidt M Volchkov V. Akpo J. Bachem O. Schlkopf B. Bauer S Gondal, M. W. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *arXiv preprint*, 2019.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230, 2018. URL http://arxiv.org/abs/1812.02230.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.