Progressive Disentanglement Using Relevant Factor VAE

Iretiayo Akinola IAKINOLA@CS.COLUMBIA.EDU
Department of Computer Science, Columbia University

Simone Fobi SF2786@COLUMBIA.EDU
Department of Mechanical Engineering, Columbia University

*

Editor: Editor's name

Abstract

Unsupervised learning of meaningful disentanglement remains an open challenge. This problem has roughly two perpendicular objectives: dimensionality reduction of high-dimensional data (such as images) to a low-dimensional latent space, and enforcing a disentanglement structure on the obtained latent space. It has been shown that improved performance in one can potentially hurt the other i.e. increased disentanglement can reduce reconstruction quality. Previous works have developed various reformulations to better decouple these objectives but there is always still a connection which often requires hyperparameter search / tuning to find the right trade-off. In this work, we propose a systematic approach that automatically adapts the relative weights of both components to obtain the right trade-off. Based on the Factor VAE approach, our method can adaptively increase or decrease the weight of disentanglement objective as a function of the discriminator performance. This makes the unsupervised learning process insensitive to the initial choice of hyperparameters and dataset-agnostic. Our approach also enables a learning curriculum that initially places focus on the reconstruction and adaptively shifts emphasis to learning disentanglements.

Keywords: disentangled representation, variational autoencoder, Factor-VAE

1. Introduction

Much of machine learning has focused on learning relevant transformations which encourage meaningful representations of the data for the task at hand. Disentanglement learning stands in contrast to these methods by proposing approaches to learn relevant factors of variations within the data which gives useful information about the structure of the data. This learned low-dimensional representation of the data presents meaningful features that can be use for subsequent tasks. Despite significant progress being made in the field of disentanglement learning, a major tradeoff between disentanglement and image reconstruction still persist. Higgins (2017) proposed beta Variational AutoEncoders (beta-VAE) which provide an unsupervised approached to learning disentangled generative factors. The beta hyper-parameter encourages more factorized representations of the latent variables, but this is done at the cost of reconstruction quality. Factor VAE (Kim and Mnih (2018)), an augmentation to beta-VAE attempts to minimize the tradeoff between disentanglement and reconstruction by introducing a discriminator which to minimize the dependence across di-

* Thanks for Aloysius Fobi for providing GPU resources to participate in this competition.

mension of the latent space and prevents penalizing total correlation unnecessarily in order to retain adequate information in the latent representation. Relevance Factor VAE (Kim et al. (2019)) builds on this approach to introduce the notion of meaningful and nuisance factors. Nuisance factors often lead to considerable drop in disentanglement performance. Thus the key idea is to learn relevance-indicator variables, which are used to drive the total correlation loss, thereby achieving superior performance in multiple metrics.

Although strong disentanglement may be achieved, poor image reconstruction suggests that the latent factors contain little information about the data of interest. Ideally, a representation which maximizes both parameters, would hold the most disentangled information about the data. Our approach, thus proposes a way to encourage both good image reconstruction and a disentangled representation. We leverage the gains made through relevance factor VAE and propose an approach to adjust the relative weights of reconstruction and disentanglement, such that the learning process is less susceptible to hyper-parameter choice. Thus our unique contributions are:

1. Leveraging discriminator performance to tune relative weights between reconstruction and disentanglement to better maximize both objectives.
2. Improving model robustness to hyper-parameter choice.
3. Adaptive method of shifting relative importance of reconstruction versus disentanglement during training.

2. Background

Variational autoencoder (vae) is a competitive generative approach to unsupervised learning from which most state-of-art approaches for disentangled learning are derived. Equation (1) gives a general version of the vae objective, commonly referred to as ELBO (evidence lower bound).

$$\mathbb{E}_{q_\phi(z|x)}\left[\mathbb{E}_{p(x)}[\log p_\theta(x|z)] - \beta \mathcal{D}_{KL}(q_\phi(z|x)||p(z))\right] \tag{1}$$

where $P(z)$ is a selected prior distribution, $q_\phi(z|x)$ is the encoder and $p_\theta(x|z)$ is the decoder. The encoder and decoder are neural network functions that are learned from this objective. The vanilla vae has $\beta = 1$.

The first term of ELBO is the reconstruction objective while the second term is the regularizer that imposes structure on the latent space, making the latent state distribution similar to the specified prior. In disentanglement learning, most methods focus on adjusting the scale and form of the second term to achieve less correlation between the dimensions of the latent space.

2.1. Factor VAE

In this work, we consider the factorization approach which decomposes the second term into two subcomponents: mutual information and total correlation. So equation (1) becomes:

$$\mathbb{E}_{q_\phi(z|x)}\left[\mathbb{E}_{p(x)}[\log p_\theta(x|z)] - I(x;z) - \beta \mathcal{D}_{KL}(q(z)||p(z))\right] \tag{2}$$

Kim and Mnih (2018) showed that this reformulation (2) prevents loss in reconstruction quality during disentanglement learning. The last term can be approximated by a discriminator that measures the level of disentanglement in the latent space. The discriminator tries to distinguish between the latent space obtained from passing the data through the encoder, and a shuffled version. The key idea is that if the disentanglement is good, the discriminator should struggle to identify the shuffled encoding as fake; shuffling the underlying factor of variation would still result in point within the data distribution.

We adopt a modified version of Factor-VAE called *Relevance Factor-VAE* Kim et al. (2019). It is a recent follow-up work that also learns to streamline redundant latent space dimension to a minimum– closer to the number of underlying factors of variation in the data.

3. Progressive Disentanglement for Factor VAEs

This work focuses on an automatic way to tune the regularization weight β in equation 2 that trades off reconstruction quality and level of disentanglement. To achieve this, we make β a function of the discriminator performance (eq(3)). This approach makes learning behavior robust to the initial choice of parameter beta β and it also makes the parameter choices dataset-agnostic. Previous work Mathieu et al. (2016) proposed an adversarial training method to disentanglement learning. In contrast to theirs, our approach is fully unsupervised; we do not use any knowledge of class/factor labels.

$$\beta' = \begin{cases} \beta * scale & \text{if discriminator accuracy is high} \\ \beta / scale & \text{otherwise} \end{cases} \tag{3}$$

scale is a scale constant that determines how much to adapt β at each update step. We found that $scale = 0.9, 0.99, 0.999$ all work well.

During training, we measure the current performance of the discriminator and determine whether to increase/decrease the weight of the disentanglement objective. In the factor-vae setting, we expect that a very low beta places more emphasis on reconstruction quality and result in a high discriminator performance; since less effort is spent on disentanglement. So if the discriminator accuracy is high (e.g. > 0.90), we increase β; conversely if the discriminator accuracy is low (e.g. < 0.55), we decrease β to focus more on the reconstruction quality. Implementation-wise, we set a single target discrimination (e.g. 0.75) and a window around it (0.05) such that the acceptable accuracy is in this range 0.75 ± 0.05 and we apply (3) whenever it is outside this range. We found that this makes the algorithm achieve stable and smooth adaptive behaviour. Another variation we implemented sets the target accuracy high (0.85) early in training and then gradually anneals it to a small value (0.60). This has the effect of focusing on reconstruction quality early on and gradually improving disentanglement level. Preliminary analysis shows this has similar performance to the single target accuracy version.

Our proposed approach achieved a top ten ranking in the disentanglement challenge; the codes for the proposed method are available here: https://gitlab.aicrowd.com/iretiayo_akinola/neurips2019_disentanglement_challenge_starter_kit.

References

Pal Burgess Glorot Botvinick Mohamed Lerchner Higgins, Matthey. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR 2.5*, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Relevance factor vae: Learning and identifying disentangled factors. *arXiv preprint arXiv:1902.01568*, 2019.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016.