Quantifying the Effects of Enforcing Disentanglement on Variational Autoencoders

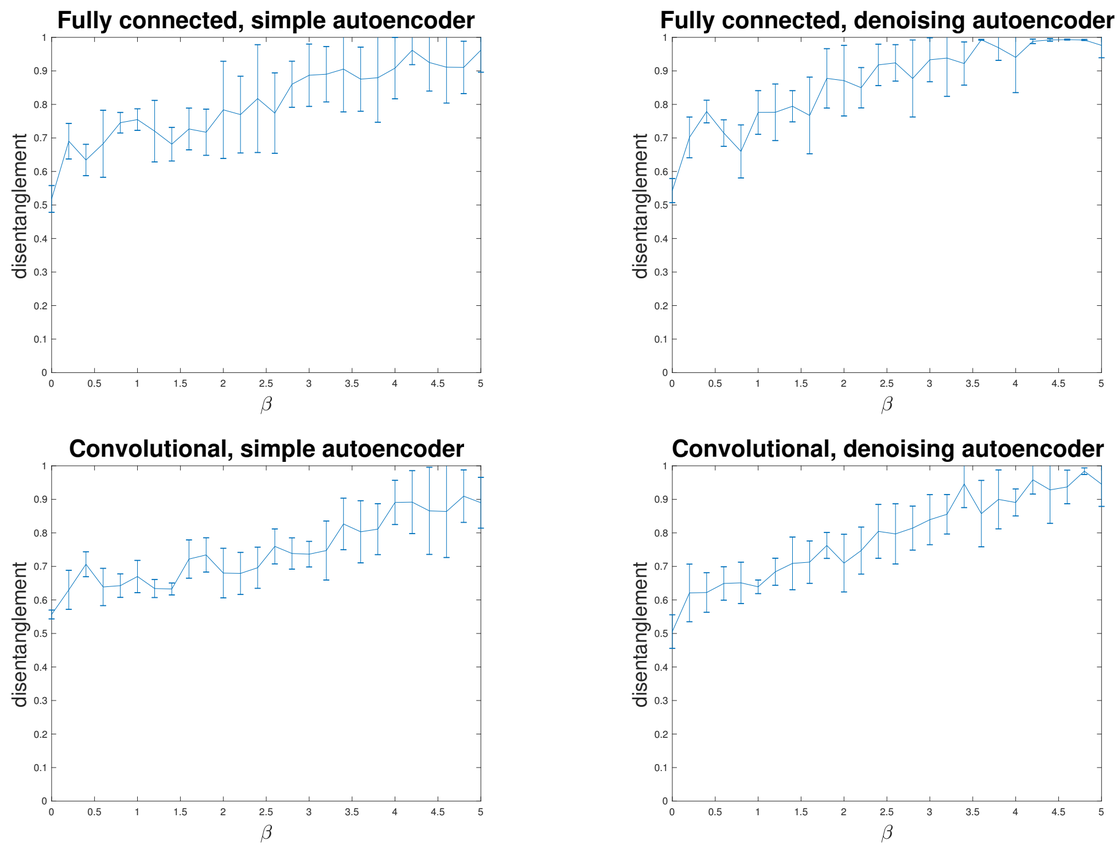
Momchil Peychev, Petar Veličković, Pietro Liò
University of Cambridge, Department of Computer Science and Technology
mpeychev@cantab.net, {petar.velickovic, pietro.lio}@cst.cam.ac.uk

Abstract

The notion of disentangled autoencoders was proposed as an extension to the variational autoencoder by introducing a disentanglement parameter β, controlling the learning pressure put on the possible underlying latent representations. For certain values of β this kind of autoencoders is capable of encoding independent input generative factors in separate elements of the code, leading to a more interpretable and predictable model behaviour. In this paper we quantify the effects of the parameter β on the model performance and disentanglement. After training multiple models with the same value of β, we establish the existence of consistent variance in one of the disentanglement measures, proposed in literature. The negative consequences of the disentanglement to the autoencoder's discriminative ability are also asserted when tested against the MNIST dataset while varying the amount of examples available during training.

Measuring Disentanglement

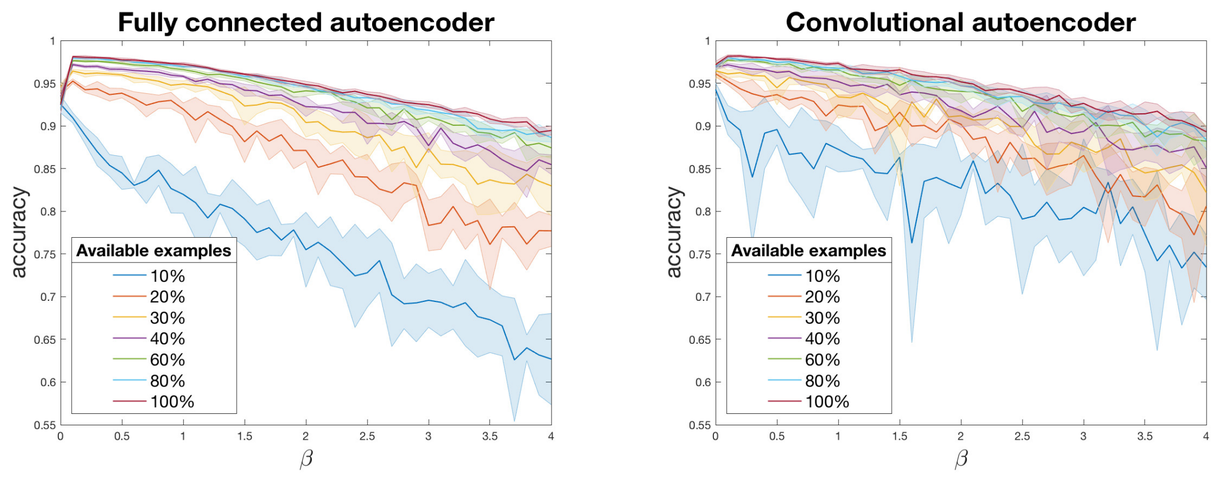
Higgins *et al.* [1] proposed a method of measuring autoencoders disentanglement. A random set of generating factors is taken, the image img_1 is constructed, and the code means $\mathbf{z}_1^\mu = encoder(img_1)$ are extracted. The same procedure is repeated, but this time one of the factors is randomly modified while all the others are kept the same. Denote the newly extracted code means with $\mathbf{z}_2^\mu$. A low capacity linear classifier is trained to map $\frac{|\mathbf{z}_1^\mu - \mathbf{z}_2^\mu|}{\max(|\mathbf{z}_1^\mu - \mathbf{z}_2^\mu|)}$ to the single factor that was changed during the process of obtaining $\mathbf{z}_1^\mu$ and $\mathbf{z}_2^\mu$. The classifier's accuracy is then reported as a disentanglement measure of the autoencoder of interest.



Disentanglement levels of the autoencoders with respect to the parameter β.

MNIST Classification with Disentangled Autoencoders

Unsupervised autoencoder is trained first and subsequently a Support Vector Machine classifier is trained (using the same training dataset) to map the image codes, produced by the encoder network, to the respective image classes. Increasing the number of training examples consistently increases the classification accuracy – providing more labels helps the model generalise. Although with higher variance, the convolutional architecture seems to be more robust to training the autoencoder with fewer datapoints. When applied to the MNIST dataset, increasing disentanglement consistently deteriorates classification accuracy. This establishes the fact that there is a trade-off between the two terms of the disentangled autoencoder loss function and that they force the model to learn different properties about the data. When training a disentangled autoencoder, this trade-off should be considered and a balanced solution is desirable.



β goes from 0 to 4 with a step of 0.1. We execute the same experiments with different number of labels available during the training.

Introduction and Background

This work primarily concerns the model of a disentangled autoencoder which represents a recent development towards building more transparent and interpretable generative models. It is capable of learning independent generating factors separately in the network, thus being more predictable in its behaviour. Given certain input data we might know what values to expect for the code and, conversely, small disturbances of the code result in expected changes of the output. We study the properties of this model with respect to changing the values of the disentanglement parameter β, measuring both its disentanglement level and discriminative ability. The implemented experiments build on top of the work of Kingma and Welling [2] on variational autoencoders and Higgins *et al.* [1, 3] on disentangled autoencoders. The latter defined the disentangled autoencoder framework by specifying the optimisation problem

$$(\phi, \theta) = \max_{\phi,\theta} \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \quad \text{subject to} \quad D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) < \epsilon \tag{1}$$

for $\epsilon > 0$. Applying the Karush-Kuhn-Tucker conditions, Equation (1) can be written as a Lagrangian

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) \tag{2}$$

with $\beta \geq 0$, deriving the final form of the cost function. A practical choice is to set $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ and $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. In this way not only the D_{KL} term can be evaluated analytically [2], but choosing $p_\theta(\mathbf{z})$ to be the isotropic normal distribution with perfectly uncorrelated components forces the model to learn representations which encode statistically independent features about the data separately, in different positions of the code. Varying the value for β regulates the amount of the applied learning pressure.

Synthetic Dataset

The observed data should possess transform continuities in order to be able to find some regularity in it in an unsupervised manner [1]. A synthetic dataset of 64x64 binarised images containing each a single shape was constructed. The generative factors defining each image are: a shape – □, ⬭ or △; position X (16 values); position Y (16 values); scale (6 levels); rotation (60 values over the $[0, \pi]$ range). The images were randomly separated in training, validation and test sets in a ratio 70:15:15 in a stratified way. Duplicate images incidentally caused by some idempotent transformations were removed. The final dataset consists of 267,021 images.

References

[1] Irina Higgins, Loïc Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning. *CoRR*, abs/1606.05579, 2016.

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[3] Irina Higgins, Loïc Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

The source code to reproduce all of our experiments described in this work can be found at www.github.com/mpeychev/disentangled-autoencoders.