Analyzing disentanglement of visual objects in semi-supervised neural networks

Andrew David Zaharia (andrew.z@columbia.edu)[1,⋆] and **Benjamin Peters (benjamin.peters@columbia.edu)**[1,⋆]

John Cunningham (jpc2181@columbia.edu)[2]

Nikolaus Kriegeskorte (n.kriegeskorte@columbia.edu)[1,3]

[1] Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA
[2] Department of Statistics and Grossman Center, Columbia University, New York, NY 10027, USA
[3] Departments of Psychology, Neuroscience, and Electrical Engineering, Columbia University, New York, NY 10027, USA
⋆ These authors contributed equally to this work

Abstract

A fundamental goal of visual systems is to condense images into compact representations of the relevant information they contain. Ideally, these representations would consist of the independent "generative factors" that fully determine, on a semantic level, the visual input. Such a "disentangled" representation could consist of the identity of a background scene, and the identity, position, pose, and size of an object. Recent research in deep neural networks (DNNs) has focused on achieving disentangled representations, through unsupervised learning, of single objects or faces in isolation. We trained and analyzed a popular DNN model of disentanglement, the β-variational autoencoder (β-VAE) on a new dataset, containing a "foreground" white circle and "background" isotropic Gaussian. We show that the neural network autoencoder architecture we use can achieve a perfectly disentangled latent representation with supervised learning, but only achieves partial disentanglement when using the unsupervised β-VAE loss function. On our dataset, higher β values result in higher reconstruction loss and greater entanglement. We propose that further inductive bias is needed to achieve better disentanglement, such as a representation which factorizes static properties and their dynamics.

Keywords: disentanglement; unsupervised learning; deep neural network; autoencoder; object vision

Disentangled visual representations compactly and independently encode the true generative factors of the world (DiCarlo & Cox, 2007; Bengio, 2009). For object recognition, such factors could be their identity, size, rotation, and color.

Recent work on disentanglement in DNNs has focused on the unsupervised learning setting. The β-VAE was designed to learn more disentangled representations by treating the reconstruction error term in the VAE loss function as a regularizer (Higgins et al., 2017; Kingma & Welling, 2014). Here we explore to what extent this approach achieves disentanglement in a scenario where such representations are achievable with the chosen architecture.

We generated images by randomly varying the positions of a circular disc occluding and a larger isotropic Gaussian, while keeping size and intensity fixed. There are 4 generative factors for this dataset: the horizontal and vertical position of the two objects. An ideal encoder for these images that is perfectly disentangled is one with four latent variables, where each one uniquely maps to one of the four generative factors. In an entangled representation, multiple latent variables will change when varying a single generative factor.

We started with a simple, 4-layer convolutional encoder network with supervised training to ensure that a perfectly disentangled encoder is achievable. We found this encoder can extract and perfectly disentangle the generating factors.

Next, we trained a β-VAE with the same encoder network architecture and a decoder network with size-matched fully connected and deconvolutional layers in reverse order, for different β. The resulting representations are entangled, and become less informative for higher β. The level of disentanglement and reconstruction quality in β-VAE further declined with increasing β, consistent with previous predictions.

In natural experience, objects are dynamic. The generative factors determining an object's appearance are likely to remain temporally stable or vary smoothly and slowly (Wiskott & Sejnowski, 2002). Such prior knowledge, as an inductive bias, could support feature learning in biological systems. We will train and analyze an autoencoder, designed to factorize static object properties and their dynamics, and predict that it will achieve better disentanglement.

References

Bengio, Y. (2009). *Learning Deep Architectures for AI* (Vol. 2) (No. 1). doi: 10.1561/2200000006

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017). β-VAE: Learning Basic Visual Concepts With A Constrained Variational Framework. *ICLR*, 1–12. doi: 10.1177/1078087408328050

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *ICLR*, 1–14. Retrieved from `http://arxiv.org/abs/1312.6114` doi: 10.1051/0004-6361/201527329

Wiskott, L., & Sejnowski, T. (2002). Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation*, *14*(4), 715–770.