Non-synergistic VAE

Gonzalo Barrientos
Department of Computer Science
University College London
London, UK
gonzalo.ayquipa.16@ucl.ac.uk

Cristina Calnegru
Department of Computer Science
University College London
London, UK
florina.calnegru.16@ucl.ac.uk

Abstract

Learning disentangling representations of the independent factors of variations that explain the data in an unsupervised setting is still a major challenge. In the following paper we address the task of disentanglement and introduce a new state-of-the-art approach called Non-synergistic variational Autoencoder (Non-Syn VAE). Our model draws inspiration from population coding, where the notion of synergy arises when we describe the encoded information by neurons in the form of responses from the stimuli. If those responses convey more information together than separate as independent sources of encoding information,they are acting synergetically. By penalizing the synergistic information within the latents we encourage information independence and by doing that disentangle the latent factors. In addition, we qualitatively compare our model with Factor VAE.

1 Introduction

Our world is hierarchical and compositional, humans can generalise better since we use primitive concepts that allow us to create complex representations [10]. Towards the creation of truly intelligent systems, they should learn in a similar way resulting in an increase of their performance since they would capture the underlying factors of variation of the data [1, 9, 3]. According to [15], a compositional representation should create new elements from the combination of primitive concepts resulting in a infinite number of new representations. Furthermore, a disentangled representations is defined as one where single latent variables are sensitive to changes in generative factors, while being invariant to changes in other factors. [1].

2 Model

The original Variational autoencoder framework [14, 17] has been used for the task mentioned before, by modifying the original ELBO formulation [11, 13, 4]; as well as the Generative Adversarial Networks [7] by encouraging the mutual information between the latents and the output of the generator [5]. To understand our model, we need first to describe Synergy [6, 18] being a popular notion of it as how much the whole is greater than the sum of its parts. It's common to describe it with the XOR gate, since we need two independent variables to fully specified the value of the output. Our hypothesis suggest that by penalising the synergistic information we encourage the model to disentangle the factors of variation. Intuitively, this means that if two latents Z_1 and Z_2 will. Computing the multivariate synergistic information is an ongoing topic of research [18, 19, 2, 8], however we decided to use the metric defined in [8], shown in Equation 1, where A_i is a non-empty subset of $\{Z_1, Z_2, ..., Z_d\}$ and the I_{max} (second term on the RHS) is defined as the specific mutual information (MI) between each outcome $x \in X$ and the subset A_i that maximises the specific mutual information. Notably, the MI can be expressed in terms of the KL divergence.

$$S_{max}(\{Z_1, Z_2, ..., Z_d\}; X) = I(\boldsymbol{Z}; X) - \sum_{x \in X} p(X = x) \max_i I(A_i; X = x) \tag{1}$$

Preprint. Work in progress.

From [12], we know that the KL term in the ELBO loss is decomposed in $D_{KL}\big[q_\phi(z_n) \parallel p(z_n)\big] + I(x_n; z)$. If we penalise the synergy defined in Eq 1, we will be penalising the MI term which is not desirable for this task [13]. Therefore, we used only I_{max}, which means maximising the subset of latents with the most amount of MI per outcome. Since it's cumbersome to maximise and minimise the latent variables, we decided to penalise the subset of latents with the minimum specific MI (ie. A_w). It's easy to see that this new equation is still a lower bound on the log likelihood $p(x)$.

$$\mathcal{L}_{new}(\theta,\phi,x) = \underbrace{\mathbb{E}_{q_\phi(z|x)}\big[\ \log p_\theta(x|z)\big]\ - D_{KL}\big[q_\phi(z|x) \parallel p(z)\big]}_{\mathcal{L}_{elbo}} - \underbrace{\alpha D_{KL}\big[q_\phi(A_w|x) \parallel p(A_w)\big]}_{\alpha * \text{Imax}} \tag{2}$$

Algorithm 1 Non Syn VAE

Input: Observations $(x^{(i)})_{i=1}^N$, batch size m, latent dimension d, weight of synergy loss α, discount factor ω, optimiser $optim$, function get_index_greedy computes A_w per batch using a greedy policy and ω.

```
θ, φ ← Initialise VAE parameters
repeat
    x^(i) ← Random minibatch B of size m, i ∈ B
    φ, θ ← optim(∇_{θ,φ} L_elbo(θ, φ; x))                 ▷ Gradients of ELBO minibatch, see Eq.2
    x'^(i) ← Random minibatch B' of size m, i ∈ B'
    worst_index ← get_index_greedy(mu, logvar, ω)          ▷ mu,logvar ~ Encoder(x'^(i), φ)
    L_syn ← α * Imax(mu, logvar, worst_index)              ▷ See Eq.2 for Imax function
    φ ← optim(∇_φ L_syn(φ; x'^(i)))                        ▷ Gradients of Syn loss minibatch
until convergence of objective
```

3 Experiments

For disentanglement, the dataset most commonly used is the dsprites dataset [16], which consists on 2D shapes generated from independent latent factors. We used the same architecture and optimizer as Factor VAE [13]. In Figure 1 (left), we see clearly that our model disentangles the factors of variation. Likewise, on the right we see the mean activation of each active latent averaged across shapes, rotations and scales.

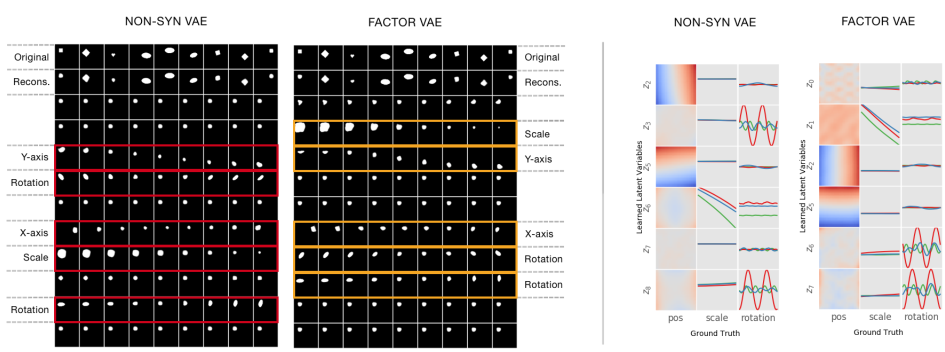

Figure 1: Left: Traverse of latents (110k steps). Right: Mean activations (110k steps)

4 Conclusions and Future work

We described a model that uses a novel approach inspired by the information theory and neuroscience fields to achieve the disentanglement of the underlying factor of variations in the data. After looking at the results,we can state that our model achieved state-of-the-art results, with a performance close to FactorVAE. As future work, we will explore other synergy metrics in the literature and will test using other datasets.

References

[1] Y. Bengio, A. C. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50. URL `https://doi.org/10.1109/TPAMI.2013.50`.

[2] N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost. Shared information – new insights and problems in decomposing information in complex systems. *CoRR*, abs/1210.5902, 2012. URL `http://arxiv.org/abs/1210.5902`.

[3] M. Botvinick, D. G. T. Barrett, P. Battaglia, N. de Freitas, D. Kumaran, J. Z. Leibo, T. Lillicrap, J. Modayil, S. Mohamed, N. C. Rabinowitz, D. J. Rezende, A. Santoro, T. Schaul, C. Summerfield, G. Wayne, T. Weber, D. Wierstra, S. Legg, and D. Hassabis. Building machines that learn and think for themselves: Commentary on lake et al., behavioral and brain sciences, 2017. *CoRR*, abs/1711.08378, 2017. URL `http://arxiv.org/abs/1711.08378`.

[4] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *CoRR*, abs/1802.04942, 2018. URL `http://arxiv.org/abs/1802.04942`.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2172–2180, 2016. URL `https://arxiv.org/abs/1606.03657`.

[6] I. Gat and N. Tishby. Synergy and redundancy among brain cells of behaving monkeys. In *Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998]*, pages 111–117, 1998. URL `http://papers.nips.cc/paper/1611-synergy-and-redundancy-among-brain-cells-of-behaving-monkeys`.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets`.

[8] V. Griffith and C. Koch. Quantifying synergistic mutual information. *CoRR*, abs/1205.4265, 2012. URL `http://arxiv.org/abs/1205.4265`.

[9] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245 – 258, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.06.011. URL `http://www.sciencedirect.com/science/article/pii/S0896627317305093`.

[10] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner. Early visual concept learning with unsupervised deep learning. *CoRR*, abs/1606.05579, 2016. URL `http://arxiv.org/abs/1606.05579`.

[11] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[12] M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

[13] H. Kim and A. Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2654–2663, 2018. URL `http://proceedings.mlr.press/v80/kim18b.html`.

[14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL `http://arxiv.org/abs/1312.6114`.

[15] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *CoRR*, abs/1604.00289, 2016. URL `http://arxiv.org/abs/1604.00289`.

[16] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[17] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1278–1286, 2014. URL `http://jmlr.org/proceedings/papers/v32/rezende14.html`.

[18] E. Schneidman, W. Bialek, and M. J. Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553, 2003. doi: 10.1523/JNEUROSCI.23-37-11539.2003. URL `http://www.jneurosci.org/content/23/37/11539`.

[19] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. URL `http://arxiv.org/abs/1004.2515`.