ISOLATING SOURCES OF DISENTANGLEMENT IN VARIATIONAL AUTOENCODERS

Tian Qi Chen, Xuechen Li, Roger Grosse, David Duvenaud

University of Toronto. Vector Institute.

CONTRIBUTIONS
Variational autoencoders naturally discover disentangled representations. To understand this behavior, we explore a refined decomposition of the KL regularization term in VAEs. We can amplify the source of disentanglement in VAEs which results in an improved algorithm with the same number of hyperparameters as the β-VAE. We call it β-TCVAE. Quantifying disentanglement is hard, and existing approaches are mostly ad hoc. We design a new measure rooted in information theory.

BACKGROUND
The penalized VAE objective can be written using the evidence lower bound (ELBO):
\[
\frac{1}{N} \sum_{n=1}^{N} \left[ \mathbb{E}_{q(z|x_n)}[\log p(x_n|z)] - \beta \text{KL}(q(z|x_n)\|p(z)) \right]
\]
\[\cdot \beta = 1 \rightarrow \text{Standard VAE objective.}
\[\cdot \beta > 1 \rightarrow \beta\text{-VAE} [1] \text{ for disentangling. Reliable in practice but not explicitly analyzed.}

NOTION OF DISENTANGLEMENT
Each dimension of a disentangled representation should:
1. Represent a different factor of variation in the data.
2. Be able to be changed independently of the other dimensions.
It is conjectured the following may be important:
1. Mutual information between the latent variables and the data.
2. Independence between the latent variables.

\[p(n) = \frac{1}{N} \quad q(z) = \frac{1}{\mathbb{E}_{q(z|x_n)}[\log p(n|z)]} \quad q(z|n) = \frac{1}{\mathbb{E}_{q(z|x_n)}[\log p(n|z)]} \quad q(z|n)p(n)
\]

DECOMPOSITION BREAKDOWN
The ELBO objective decreases all three terms:
1. Mutual information between the training data and the latent variables [2].
2. Total correlation (TC) between the latent variables. A measure of statistical dependence.
3. Dimension-wise KL. Simple regularization acting on each dimension of the representation.

MINIBATCH-BASED ESTIMATION
We can train with arbitrary weights on each term if we can stochastically estimate \(\log q(z)\) and \(\log q(z|n)\).

Problem. Evaluation of \(q(z)\) depends on full data.
Solution. Estimate \(q(z)\) based on the current mini-batch, and weight appropriately. Inspired by importance sampling:
\[\mathbb{E}_{q(z)}[\log q(z)] \approx \frac{1}{M} \sum_{i=1}^{M} \log \frac{1}{N} \sum_{n=1}^{N} q(z|n) p(n)\]
where \(z|n\) is a sample from \(q(n)\)

SPECIAL CASE: β-TCVAE
We designate a special case of the decomposition as a meaningful algorithm for learning disentangled representations, the β-TCVAE objective:
\[
\frac{1}{N} \sum_{n=1}^{N} \left[ \mathbb{E}_{q(z|x_n)}[\log p(n|z)] \right] = \text{Index Code MI} + \text{Total Correlation} + \text{Dimension-wise KL}
\]

MEASURING DISENTANGLEMENT
If we have a set of latent variables \(\{z_i\}\) and set of known factors \(\{v_k\}\), then we can use the empirical mutual information \(I_{\beta}(z_i; v_k)\) to quantify how well a latent variable \(z_i\) reflects a ground truth factor \(v_k\). The full metric we call mutual information gap (MIG) is
\[
\frac{1}{K} \sum_{k=1}^{K} \frac{1}{N} \sum_{v_k} I_{\beta}(z_i; v_k) - \max_{j} I_{\beta}(z_i; v_j)
\]
where \(j^{(k)} = \arg\max_j I_{\beta}(z_i; v_j)\) and \(K\) is the number of known factors.

The gap encourages two important properties:
- Axis-alignment of the representation.
- Compactness of the representation.

QUALITATIVE RESULTS

DISENTANGLED VS. INDEPENDENT REPRESENTATIONS

QUANTITATIVE COMPARISONS

REFERENCES