B-VAE : Learning Basic Visual Concepts With A Constrained Variational Framework

2017, ICLR, Irina Higgins et. al

cf. **disentangled representation** can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors (Bengio et al., 2013)

간단하게 VAE에서 ELBO Term의 KL 부분 앞에 B를 도입함

Contribution

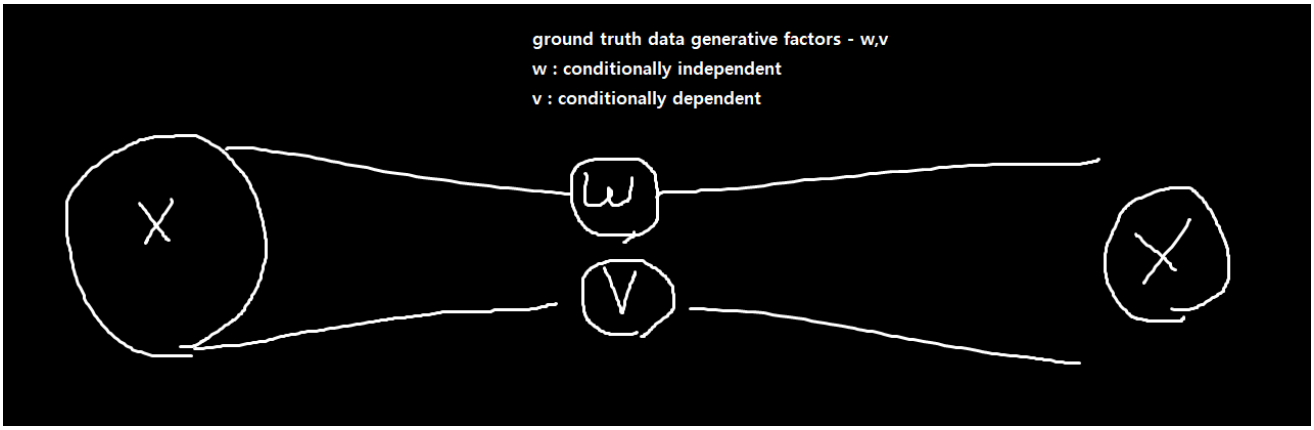
1) we proposed B-VAE, a new unsupervised approach for learning disentangled representations of independent visual factors

2) we devise a protocol to quantitatively compare the degree of disentanglement learnt by different models

3) 우리의 것이 sota !

Model

Let $\mathcal{D} = \{X, V, W\}$ be the set that consists of images $\mathbf{x} \in \mathbb{R}^N$ and two sets of ground truth data generative factors: conditionally independent factors $\mathbf{v} \in \mathbb{R}^K$, where $\log p(\mathbf{v}|\mathbf{x}) = \sum_k \log p(v_k|\mathbf{x})$; and conditionally dependent factors $\mathbf{w} \in \mathbb{R}^H$. We assume that the images $\mathbf{x}$ are generated by the true world simulator using the corresponding ground truth data generative factors: $p(\mathbf{x}|\mathbf{v}, \mathbf{w}) = \mathbf{Sim}(\mathbf{v}, \mathbf{w})$.



아래에서 식 2를 보게 된다면, 쉽게 이야기해서 ELBO Term에서 KL divergence가 e 미만이 되길 원한다는 식으로 생각.

이 때 KL divergence Term은 approximate posterior distribution이 prior (isotropic unit Gaussian distribution)을 따르고 이 말은 latent variable이 generative factors v를 disentangled manner로 capture하기를 원한다는 것이다. (쉽게 말해서 독립성을 따르는 ground truth generative factor, w를 latent variable z가 capture 하게끔 하기 위해서 KL divergence term을 둔다는 것임.)

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \quad \text{subject to } D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon \tag{2}$$

Re-writing Eq. 2 as a Lagrangian under the KKT conditions (Kuhn & Tucker, 1951; Karush, 1939), we obtain:

$$\mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\left(D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - \epsilon\right) \tag{3}$$

where the KKT multiplier β is the regularisation coefficient that constrains the capacity of the latent information channel $\mathbf{z}$ and puts implicit independence pressure on the learnt posterior due to the isotropic nature of the Gaussian prior $p(\mathbf{z})$. Since $\beta, \epsilon \geq 0$ according to the complementary slackness KKT condition, Eq. 3 can be re-written to arrive at the β-VAE formulation - as the familiar variational free energy objective function as described by Jordan et al. (1999), but with the addition of the β coefficient:

$$\mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) \geq \mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{z},\beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{4}$$

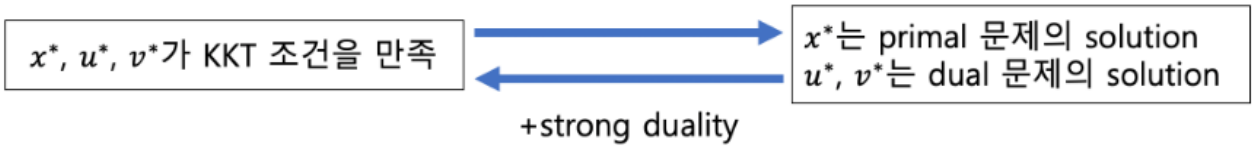
cf. KKT condition

Given general problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0,\ i = 1,\dots m \\ & \ell_j(x) = 0,\ j = 1,\dots r \end{aligned}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial\left(f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)\right)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \leq 0,\ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i (dual feasibility)



KKT condition을 만족한다고 했으니, 충분 조건으로 생각하면 된다.

We postulate that in order to learn disentangled representations of the conditionally independent data generative factors v, it is important to set β > 1, thus putting a stronger constraint on the latent bottleneck than in the original VAE formulation of Kingma & Welling (2014).

즉, conditionally independent generative factor를 담아내기 위해서 β 를 1보다 크게 하는 것임.

we hypothesis that higher values of β should encourage learning a disentangled representation of v.

즉, β를 크게 할수록 disentangled representation을 더 잘 학습한다고 가정함.

그러나 적절한 β를 찾는 것이 쉽지는 않다고 한다. (추가적으로 공부할 부분)

Disentanglement Metric

As stated above, we assume that the data is generated by a ground truth simulation process which uses a number of data generative factors, some of which are **conditionally independent**, and we also assume that they are **interpretable**.

PCA, ICA 등 역시도 representation leaning을 하게 되면 conditionally independent한 variables를 만들 수 있음. 그러나 그러한 variables은 intepretable하다고 할 수 없다.

그래서 우리가 제안하는 것은, intepretablility & independence를 measure하는 것임.

절차를 살펴보자

이 때 y(position x, position y, scale, rotation) 은 v(independent generative factors)라고 하자.

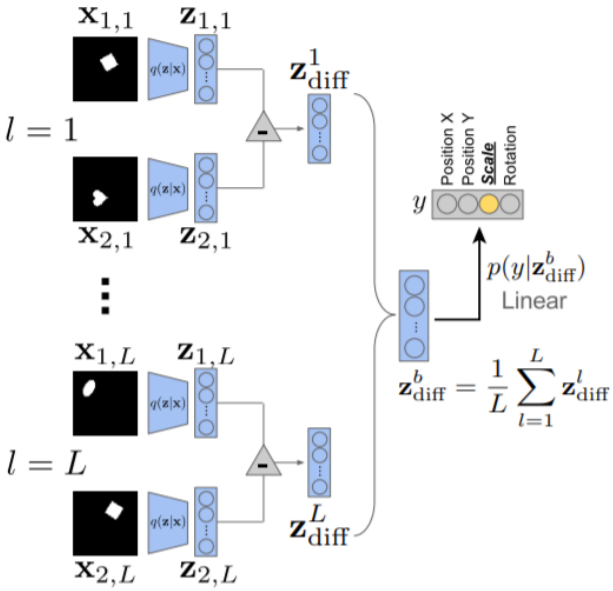


Figure 5: Schematic of the proposed disentanglement metric: over a batch of L samples, each pair of images has a fixed value for one target generative factor y (here $y = scale$) and differs on all others. A linear classifier is then trained to identify the target factor using the average pairwise difference $\mathbf{z}^b_{\text{diff}}$ in the latent space over L samples.

More formally, we start from a dataset $\mathcal{D} = \{X, V, W\}$ as described in Sec. 2, assumed to contain a balanced distribution of ground truth factors $(\mathbf{v}, \mathbf{w})$, where images data points are obtained using a ground truth simulator process $\mathbf{x} \sim \mathbf{Sim}(\mathbf{v}, \mathbf{w})$. We also assume we are given labels identifying a subset of the independent data generative factors $\mathbf{v} \in V$ for at least some instances.
We then construct a batch of B vectors $\mathbf{z}^b_{\text{diff}}$, to be fed as inputs to a linear classifier as follows:

1. Choose a factor $y \sim Unif[1...K]$ (e.g. $y = scale$ in Fig. 5).

2. For a batch of L samples:
   (a) Sample two sets of latent representations, $\mathbf{v}_{1,l}$ and $\mathbf{v}_{2,l}$, enforcing $[\mathbf{v}_{1,l}]_k = [\mathbf{v}_{2,l}]_k$ if $k = y$ (so that the value of factor $k = y$ is kept *fixed*).
   (b) Simulate image $\mathbf{x}_{1,l} \sim \mathbf{Sim}(\mathbf{v_{1,l}})$, then infer $\mathbf{z}_{1,l} = \mu(\mathbf{x}_{1,l})$, using the encoder $q(\mathbf{z}|\mathbf{x}) \sim N(\mu(\mathbf{x}), \sigma(\mathbf{x}))$.
   Repeat the process for $\mathbf{v}_{2,l}$.
   (c) Compute the difference $\mathbf{z}^l_{\text{diff}} = |\mathbf{z}_{1,l} - \mathbf{z}_{2,l}|$, the absolute linear difference between the inferred latent representations.

3. Use the average $\mathbf{z}^b_{\text{diff}} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{z}^l_{\text{diff}}$ to predict $p(y|\mathbf{z}^b_{\text{diff}})$ (again, $y = scale$ in Fig. 5) and report the accuracy of this predictor as **disentangement metric score**.

The classifier's goal is to predict the index y of the generative factor that was kept fixed for a given $\mathbf{z}^b_{\text{diff}}$. The accuracy of this classifier over multiple batches is used as our disentanglement metric score. We choose a linear classifier with low VC-dimension in order to ensure it has no capacity to perform nonlinear disentangling by itself. We take differences of two inferred latent vectors to reduce the variance in the inputs to the classifier, and to reduce the conditional dependence on the inputs $\mathbf{x}$. This ensures that on average $\left[\mathbf{z}^b_{\text{diff}}\right]_y < \left[\mathbf{z}^b_{\text{diff}}\right]_{\{\setminus y\}}$. See Equations 5 in Appendix A.4 for more details of the process.

A.4 DISENTANGLEMENT METRIC DETAILS

We used a linear classifier to learn the identity of the generative factor that produced $\mathbf{z}^b_{\text{diff}}$ (see Equations (5) for the process used to obtain samples of $\mathbf{z}^b_{\text{diff}}$). We used a fully connected linear

classifier to predict $p(y|\mathbf{z}^b_{\text{diff}})$, where y is one of four generative factors (position X, position Y, scale and rotation). We used softmax output nonlinearity and a negative log likelihood loss function. The classifier was trained using the Adagrad (Duchi et al., 2011) optimisation algorithm with learning rate of 1e-2 until convergence.

$$
\begin{aligned}
&\mathcal{D} = \{V \in \mathbb{R}^K, W \in \mathbb{R}^H, X \in \mathbb{R}^N\},\ y \sim Unif[1...K] \\
&\text{Repeat for } b = 1 \dots B: \\
&\mathbf{v}_{1,l} \sim p(\mathbf{v}),\ \mathbf{w}_{1,l} \sim p(\mathbf{w}),\ \mathbf{w}_{2,l} \sim p(\mathbf{w}),\ [\mathbf{v}_{2,l}]_k = \begin{cases} [\mathbf{v}_{1,l}]_k, & \text{if } k = y \\ \sim p(v_k), & \text{otherwise} \end{cases} \\
&\mathbf{x}_{1,l} \sim \mathbf{Sim}(\mathbf{v_{1,l}}, \mathbf{w_{1,l}}),\ \mathbf{x_{2,l}} \sim \mathbf{Sim}(\mathbf{v_{2,l}}, \mathbf{w_{2,l}}), \\
&q(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), \sigma(\mathbf{x})),\ \mathbf{z}_{1,l} = \mu(\mathbf{x}_{1,l}),\ \mathbf{z}_{2,l} = \mu(\mathbf{x}_{2,l}) \\
&\mathbf{z}^l_{\text{diff}} = |\mathbf{z}_{1,l} - \mathbf{z}_{2,l}|, \quad \mathbf{z}^b_{\text{diff}} = \frac{1}{L}\sum^{L} \mathbf{z}^l_{\text{diff}}
\end{aligned}
\tag{5}
$$

식 5에서 $[v_{2,l}]_k = [v_{1,l}]_k$ if $k = y$

즉 conditionally independent ground truth generative factors의 차원은 K인데 여기서 하나를 고정시킨다(예시에선 scale)

그런 다음에 x를 생성해내고, (z = v,w 라고 생각하면 될 듯)

이 다음에 x를 활용해서 z를 생성해낸다. (z = mu(x) 로 상정함 -- 이 부분이 나중에 다른 논문에서 문제점으로 제기됨)

그 다음에 서로 빼면 예상으론 scale에 관계 된 z의 feature는 0이 될 터이고. 이것을 단순하게 classifier만 먹여도 p(y|z_diff)는 쉽게 scale에 관계된 부분이라고 예측할 수 있겠지 (y=[0,0,1,0]으로 하면 되지 않을 까 함.)

Conclusion

Unlike InfoGAN and DC-IGN. our approach does not depend on any a priori knowledge about the number or the nature of data generative factors

<- 근데 metric을 활용하려면 개수를 알고 있어야 될 것 같네 (이것도 추후에 문제가 되겠군)

< - 즉 정말로 unsupervised learning이냐 이거지.. inductive bias가 들어갈텐데..

나의 정리

- Disentangled representation learning 임 (처음 접함)
- VAE에서 KL Term에 β를 도입해서 매우 간단하게 해서 representation learning 실시
- 측정 방식 (참신했음.)
- 허나, 정말로 unsupervised 하냐에 대해서는 물음표