35th AAAI Conference on Artificial Intelligence
February 2-9, 2021

SHANGHAI JIAO TONG UNIVERSITY

Disentangled Information Bottleneck

Ziqi Pan, Li Niu, Jianfu Zhang, Liqing Zhang

MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University

Contents

- The trade-off problem in IB Lagrangian
  - The information bottleneck method & the IB Lagrangian
  - The trade-off problem
- Our method
  - Maximum compression
  - Consistency property on maximum compression
  - Our objective function from the perspective of supervised disentangling
- Experiments
  - Information compression
  - Supervised disentangling

The IB Lagrangian Trade-off

Theorem 1. *Consider the derivable IB Lagrangian,*

$$\mathcal{L}_{\mathrm{IB}}\left[q\left(T|X\right);\beta\right] = -I\left(T;Y\right) + \beta I\left(X;T\right),$$

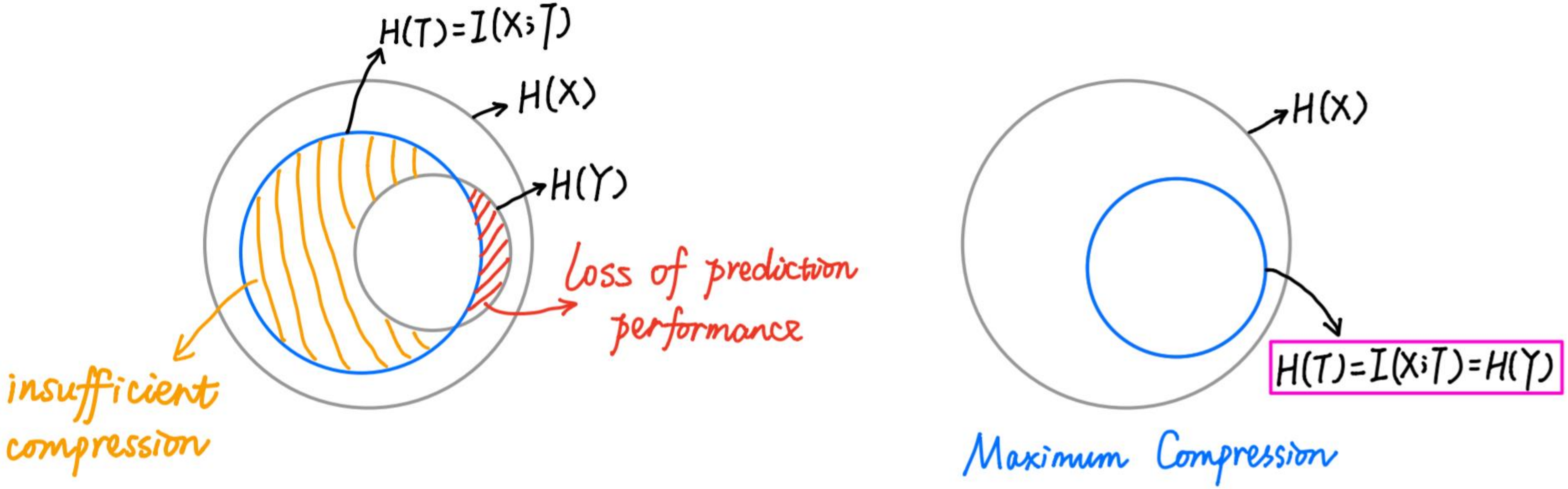
to be minimized over q with $\beta \geqslant 0$. Let q^*_β optimize $\mathcal{L}_{\mathrm{IB}}\left[q\left(T|X\right);\beta\right]$. Assume that $I_{q^*_\beta}\left(X;T\right) \neq 0$,

$$\frac{\partial I_{q^*_\beta}\left(T;Y\right)}{\partial\beta} < 0 \text{ and } \frac{\partial I_{q^*_\beta}\left(X;T\right)}{\partial\beta} < 0.$$

- For every nontrivial solution q^*_β such that $I_{q^*_\beta}(X;T) \neq 0$, $I(T;Y)$ strictly decreases as β increases.
- In fact, the proof is completed by changing probabilistic mapping $q(T|X)$ towards the aggregated distribution $q(T) = \frac{1}{n}\sum_{i=1}^{n} q(T|x_i)$, which strictly reduces $I(X;T)$ due to the concavity of the entropy $H(T)$.

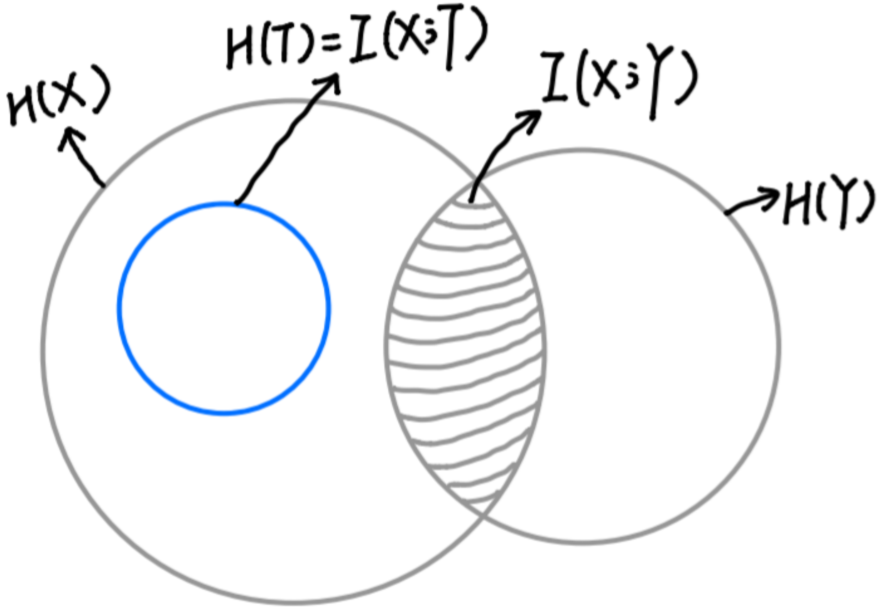
Maximum Compression

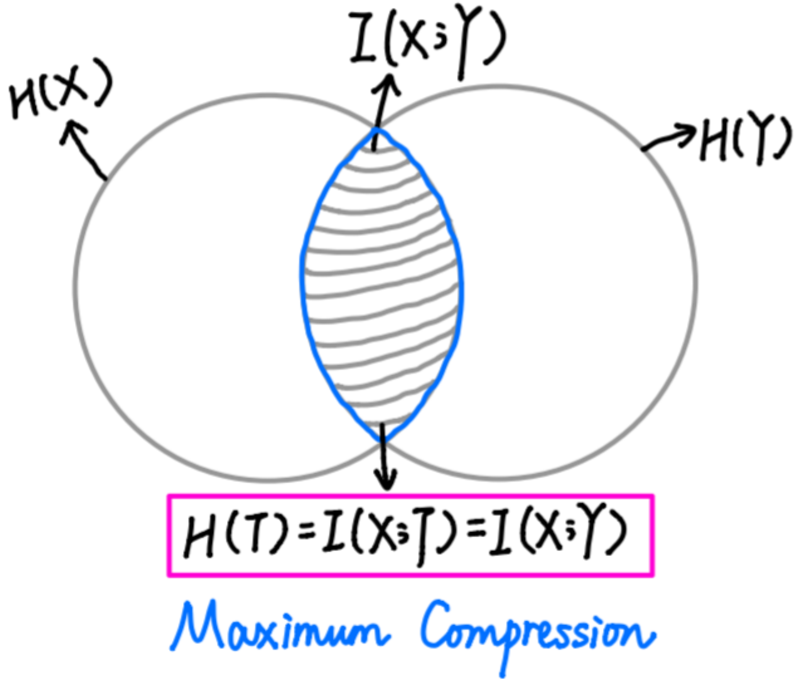
- Given source random variable X and target random variable Y, we expect to compress X maximally into T without reducing $I(T;Y)$, namely tackle the trade-off problem.
- Quantifying the maximum compression case (using Venn diagram):
  - Y is a deterministic function of X:



Maximum Compression

- Given source random variable X and target random variable Y, we expect to compress X maximally into T without reducing $I(T;Y)$, namely tackle the trade-off problem.
- Quantifying the maximum compression case (using Venn diagram):
  - Generalized case:





Consistency Property on Maximum Compression

- The maximum compression case:
$$I(X;T) = I(T;Y) = I(X;Y)$$
  - In case of Y is a deterministic function of X, $I(X;Y)$ becomes $H(Y)$.
- We aim to design a cost functional $\mathcal{L}$, such that the maximum compression case is expected to be obtained via minimizing $\mathcal{L}$.
  - Specifically, we expect that minimized $\mathcal{L}$ consistently satisfies $I(X;T) = I(T;Y) = I(X;Y)$.
- The formal definition of *consistency* on maximum compression is given as

Definition 1 (Consistency)**.** *The lower-bounded cost functional $\mathcal{L}$ is consistent on maximum compression, if*

$$\forall \epsilon > 0, \exists \delta > 0, \quad \mathcal{L} - \mathcal{L}^* < \delta \to |I(X;T) - H(Y)| + |I(T;Y) - H(Y)| < \epsilon,$$

where $\mathcal{L}^*$ is the global minimum of $\mathcal{L}$.

Our Objective Function

- After realizing the relation between IB and supervised disentangling, we implement the IB from the perspective of supervised disentangling:

$$\mathcal{L}_{\text{DisenIB}}\left[q\left(S|X\right), q\left(T|X\right)\right] = -I\left(T;Y\right) - I\left(X;S,Y\right) + I\left(S;T\right).$$

  - Encourage (S, Y) to represent the overall information of X by maximizing $I(X; S, Y)$, so that S at least covers the information of Y-irrelevant data aspect.
  - Encourage that Y can be accurately decoded from T by maximizing $I(T; Y)$, so that T at least covers the information of Y-relevant data aspect.
  - Hence, the amount of information stored in S and T are both lower bounded. In such a case, forcing S to be disentangled from T by minimizing $I(S; T)$ eliminates the overlapping information between them and thus tightens both bounds, leaving the exact information relevant (resp., irrelevant) to Y in T (resp., S).
- The maximum compression can be consistently achieved via optimizing $\mathcal{L}_{\text{DisenIB}}$.

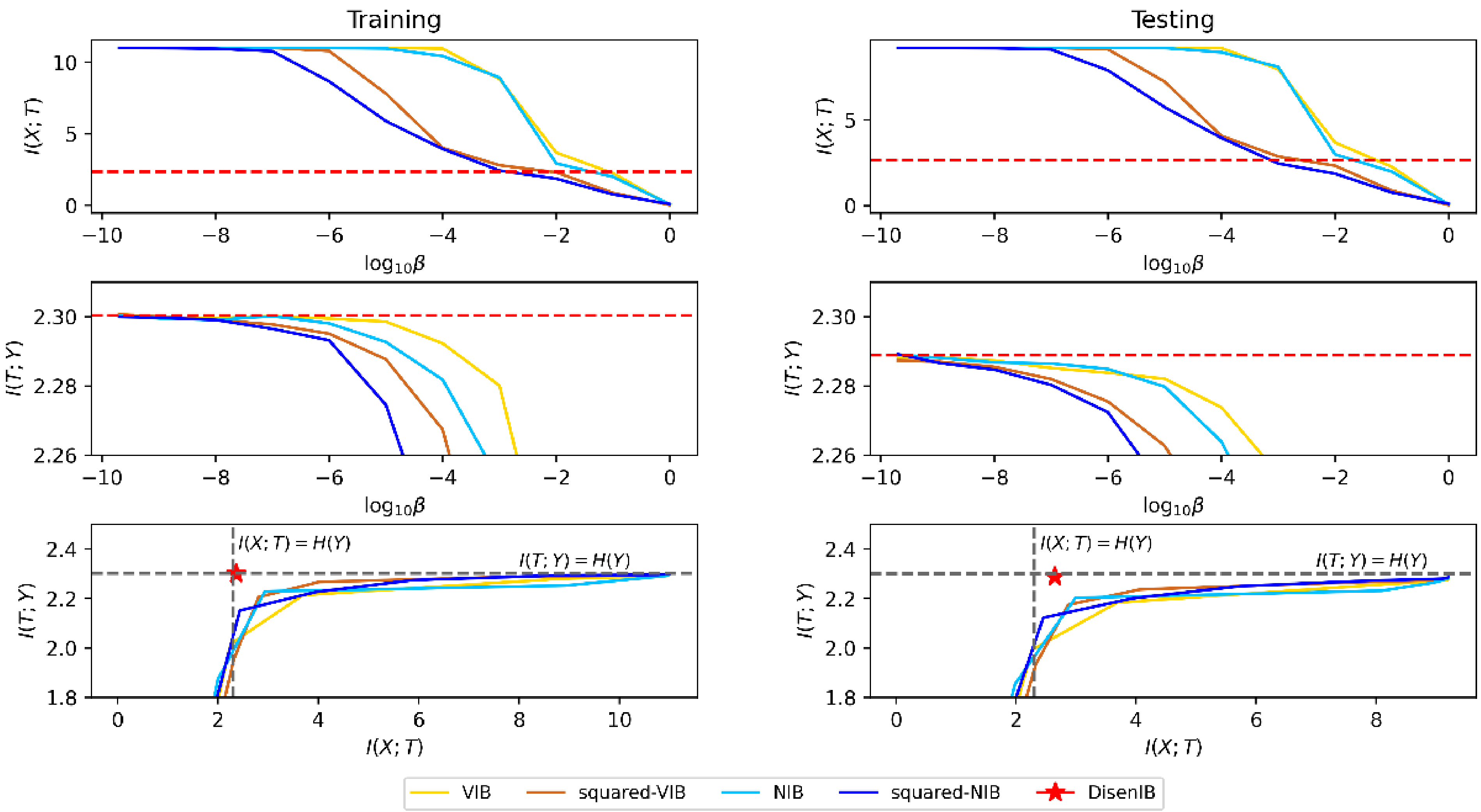
Theorem 2. $\mathcal{L}_{\text{DisenIB}}$ *is consistent on maximum compression.*

Practical Implementation

- Using variational approximations maximize $I(T;Y)$ and $I(X;S,Y)$:
  - By introducing variational probabilistic mapping $p(y|t)$ **(decoder)**:
    $$I(T;Y) \geq \mathbb{E}_{q(y,t)} \log p(y|t) + H(Y)$$
  - By introducing variational probabilistic mapping $r(x|s,y)$ **(reconstructor)**:
    $$I(X;S,Y) \geq \mathbb{E}_{q(x,s,y)} \log r(x|s,y) + H(X)$$
  - Using *density-ratio-trick* to minimizing $I(S;T)$ by involving a **discriminator** d:
    $$\min_{\mathrm{q}} \max_{d} \mathbb{E}_{q(s)q(t)} \log d(s,t) + \mathbb{E}_{q(s,t)} \log(1 - d(s,t))$$
- Code is available at https://github.com/PanZiqiAI/disentangled-information-bottleneck

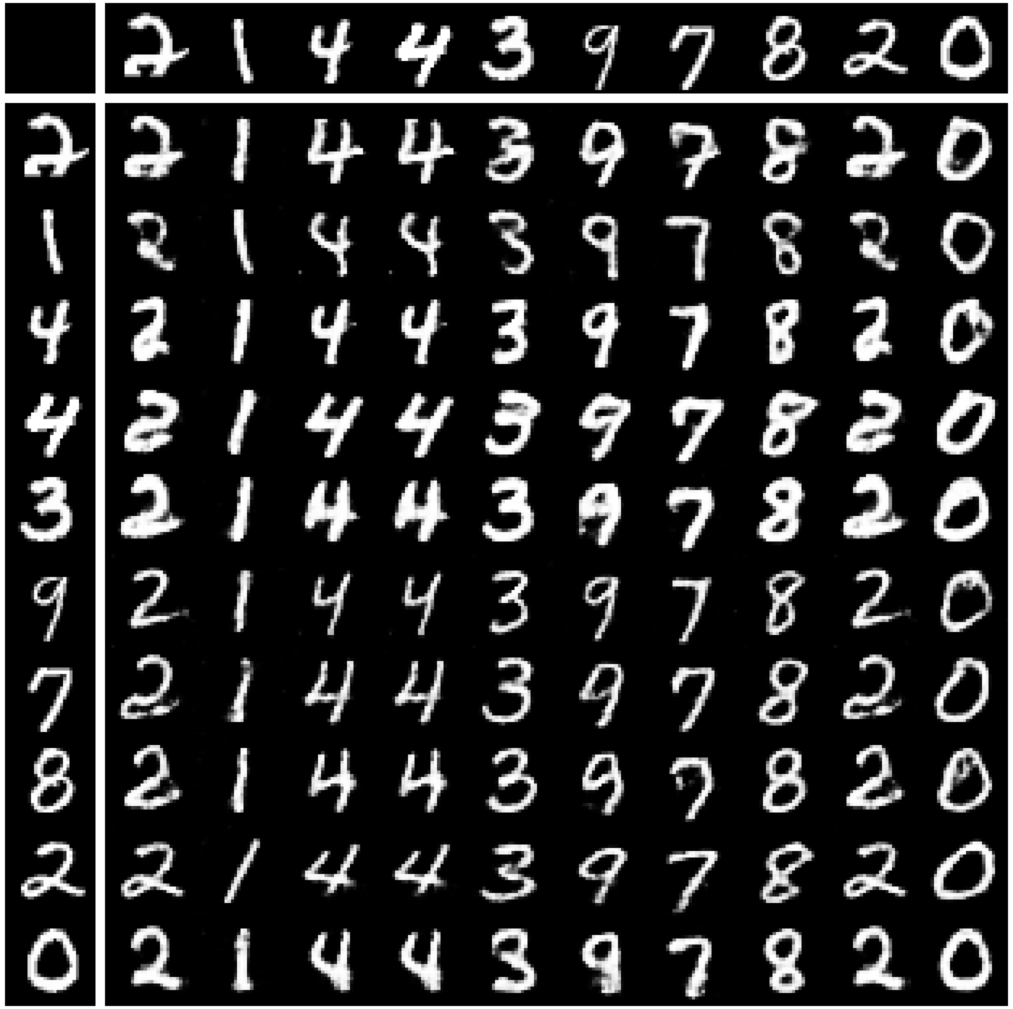
Experimental Results

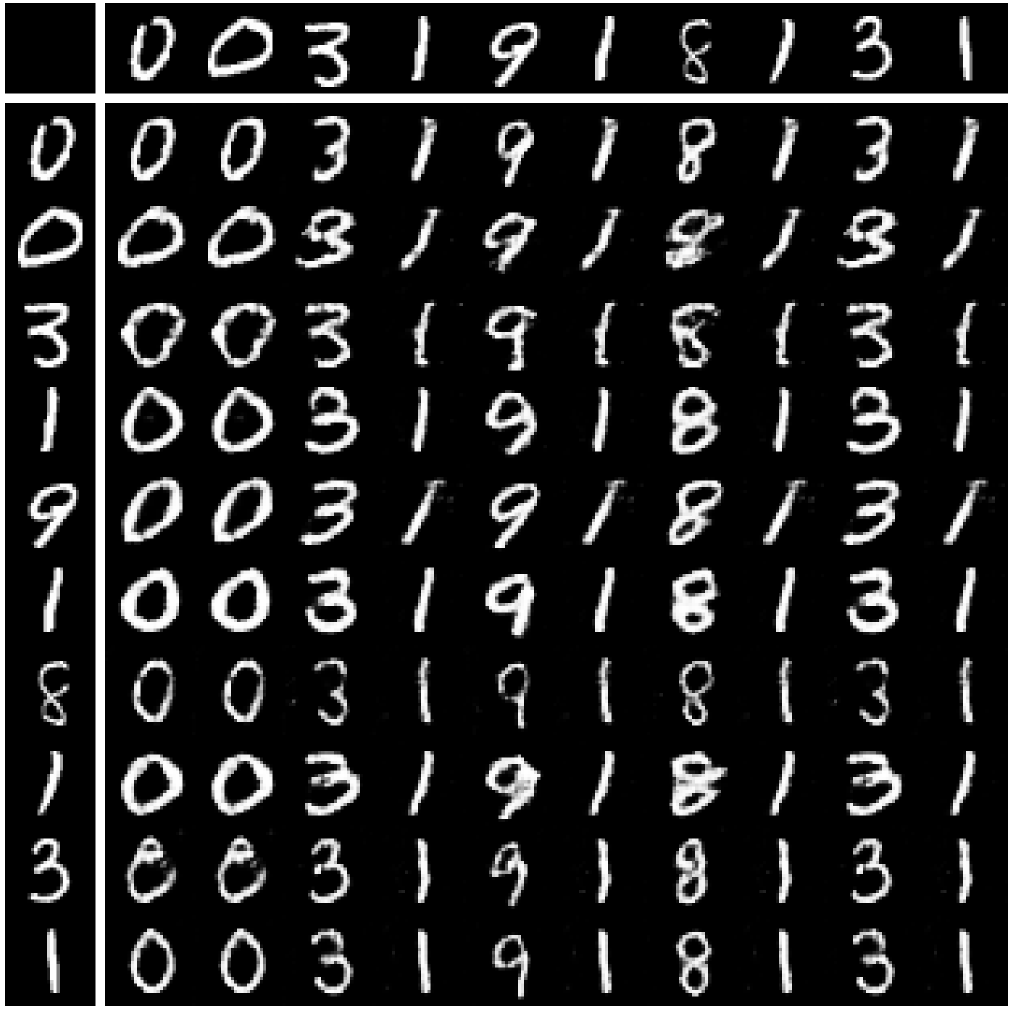
- Behavior on *IB Plane*



Experimental Results

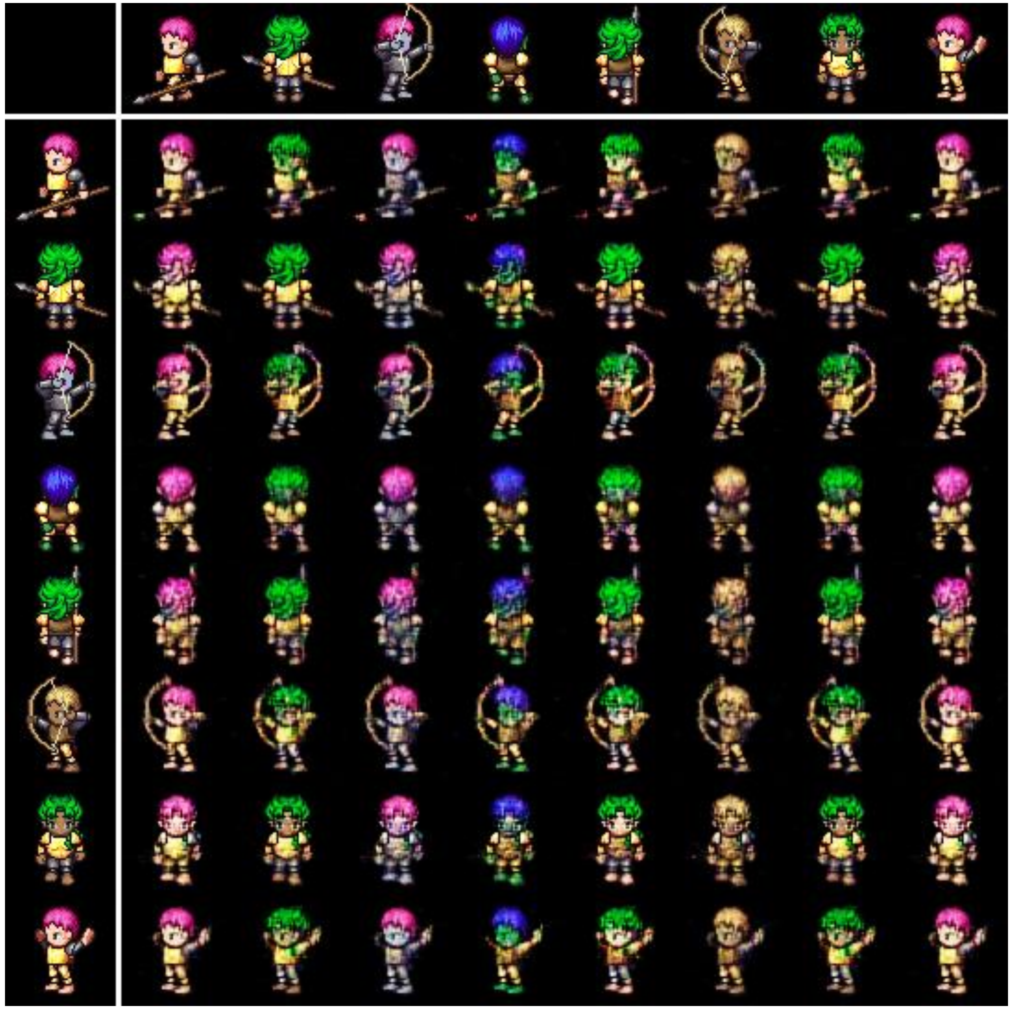
- Supervised disentangling (MNIST)





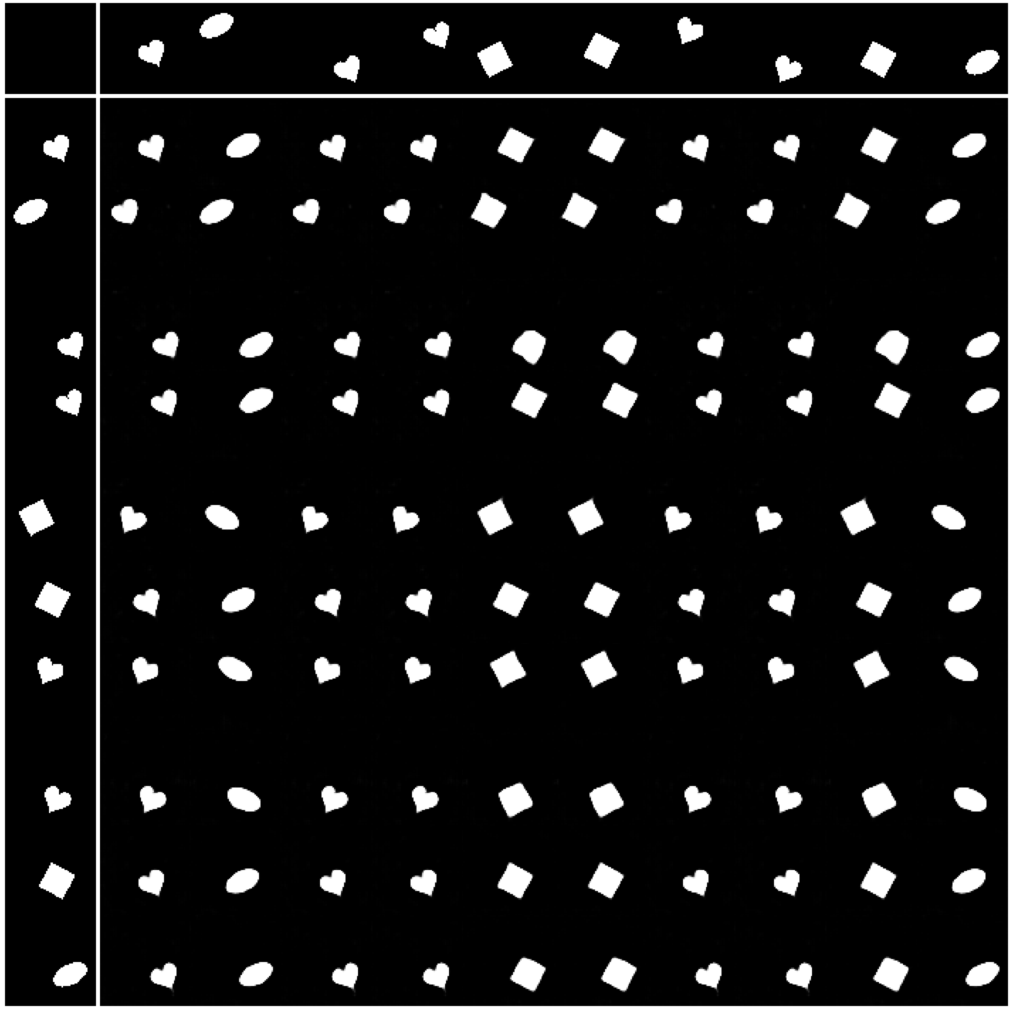
Experimental Results

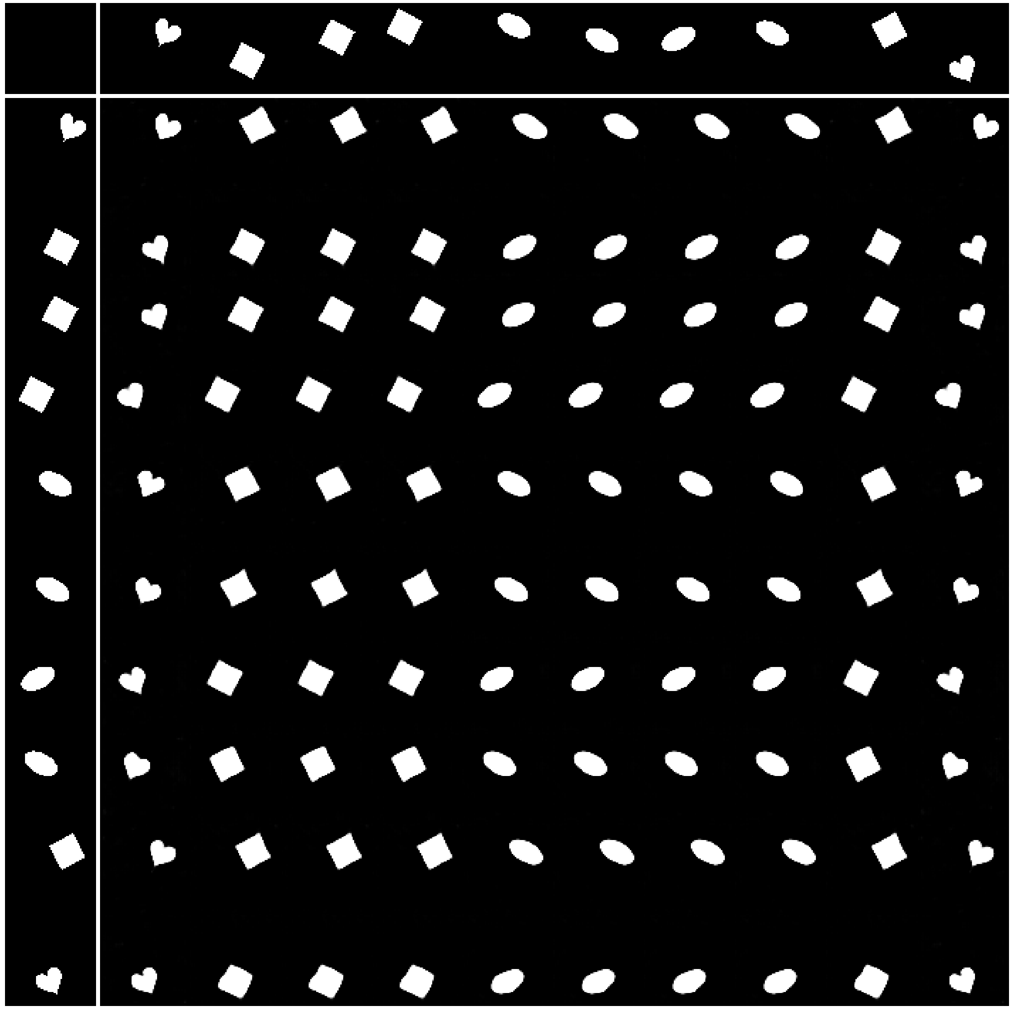
- Supervised disentangling (Sprites)



Experimental Results

- Supervised disentangling (Shapes)





Thanks for watching!

Disentangled Information Bottleneck

Ziqi Pan, Li Niu, Jianfu Zhang, Liqing Zhang

MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University