Variational Interaction Information Maximization for Cross-domain Disentanglement

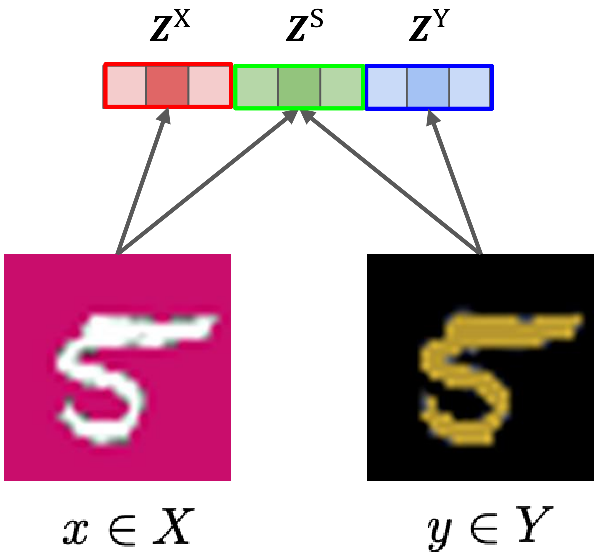
HyeongJoo Hwang, Geon-Hyeong Kim

Seunghoon Hong, Kee-Eung Kim

KAIST

Cross-domain disentanglement learning

- Given a set of paired data (x,y) sampled from unknown joint distribution p(x,y), learn a structured representation that can be factorized into three parts



- **Domain-specific representation Z^X and Z^Y**, capturing exclusive factors of variations in domain X and Y
- **Shared representation Z^S**, capturing common factors shared across domains

=> Disentangled representations gives us <u>interpretability</u> on both the data and the model.

Cross-domain disentanglement learning

- Two data domains **X**,**Y** are paired according to some shared factors of variations.

  Ex) MNIST-CDCB

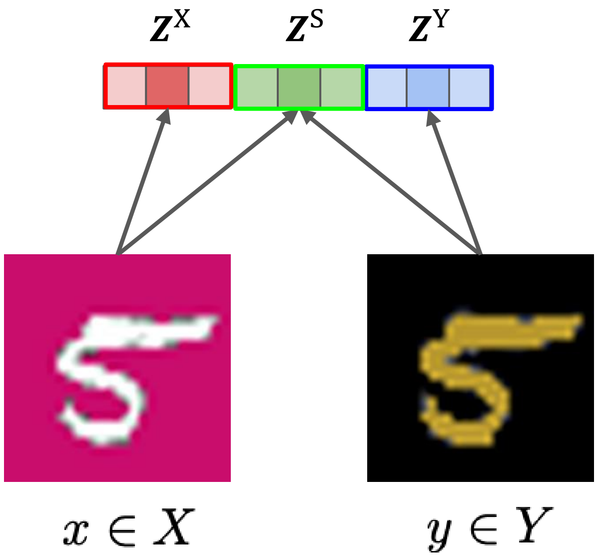
X: MNIST w/ Colored Background

Y: MNIST w/ Colored Digit

- Common factors of variation: Identity, shape, style of digits.
- Exclusive factors in X: Color of the background.
- Exclusive factors in Y: Color of the digit.

Cross-domain disentanglement learning

- Given a set of paired data (x,y) sampled from unknown joint distribution p(x,y), learn a structured representation that can be factorized into three parts



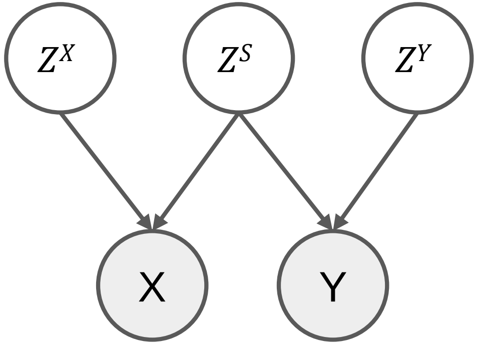
Z^X : BG color of X

Z^S : Digit ID of X and Y

Z^Y : Digit color of Y

Unsupervised cross-domain disentanglement learning

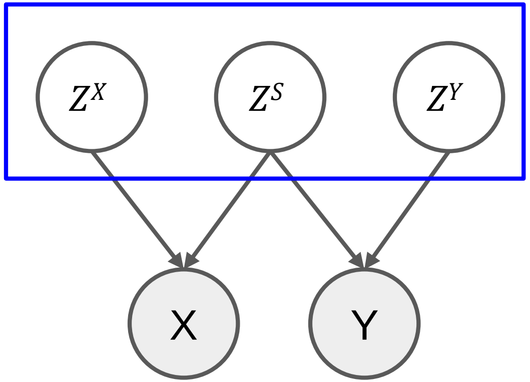
- Given a set of paired data (x,y) sampled from unknown joint distribution p(x,y), learn a structured representation that can be factorized into three parts



- **Domain-specific representation Z^X and Z^Y**, capturing exclusive factors of variations in domain X and Y
- **Shared representation Z^S**, capturing common factors shared across domains

Unsupervised cross-domain disentanglement learning

- Given a set of paired data (x,y) sampled from unknown joint distribution p(x,y), learn a structured representation that can be factorized into three parts

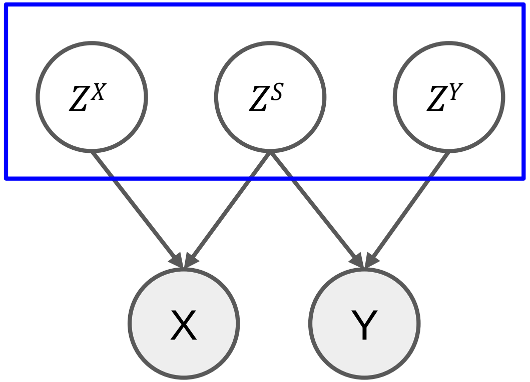


- **Domain-specific representation Z^X and Z^Y**, capturing exclusive factors of variations in domain X and Y
- **Shared representation Z^S**, capturing common factors shared across domains

Q1. How do we learn informative representation without labels?

Unsupervised cross-domain disentanglement learning

- Given a set of paired data (x,y) sampled from unknown joint distribution p(x,y), learn a structured representation that can be factorized into three parts



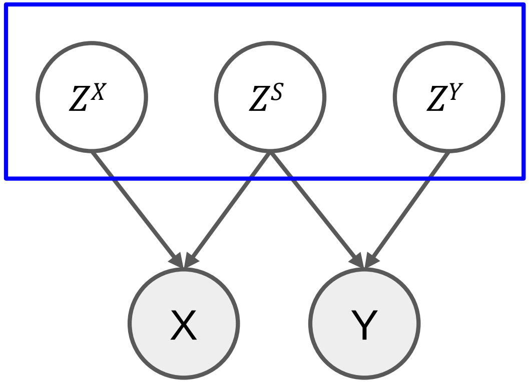
- **Domain-specific representation Z^X and Z^Y**, capturing exclusive factors of variations in domain X and Y
- **Shared representation Z^S**, capturing common factors shared across domains

Q1. How do we learn informative representation without labels?

A1. Learn a generative model to approximate p(x,y) using Z^X, Z^Y and Z^S

Unsupervised cross-domain disentanglement learning

- Given a set of paired data (x,y) sampled from unknown joint distribution p(x,y), learn a structured representation that can be factorized into three parts



- **Domain-specific representation Z^X and Z^Y**, capturing exclusive factors of variations in domain X and Y
- **Shared representation Z^S**, capturing common factors shared across domains

Q2. How do we enforce disentanglement constraints?

Enforcing factorization via regularization

- Add regularization on encoder (q) to enforce disentanglement

$$\max \mathcal{L} = \max_{p,q} \mathcal{L}_{ELBO}(p,q) + \lambda \cdot \mathcal{L}_{disentangle}(q)$$

Desiderata of cross-domain disentanglement (imposed by regularization)

1. **Decomposition** : the factors in Z^X and Z^Y should be exclusive to each domain, while all shared information is captured by Z^S

2. **Disentanglement**: the factors in Z^X, Z^Y and Z^S should be mutually exclusive

Joint regularization

$$\max_q \mathcal{L}_{disentangle}(q) = 2 \cdot \underbrace{I(X;Y;Z^S)}_{\text{Interaction Information}} - \underbrace{I(Z^X;Z^S) - I(Z^Y;Z^S)}_{\text{Mutual information(s)}}$$

Joint regularization

$$\max_q \mathcal{L}_{disentangle}(q) = 2 \cdot \underbrace{\boxed{I(X;Y;Z^S)}}\, - I(Z^X; Z^S) - I(Z^Y; Z^S)$$

Interaction information:

The amount of information shared among three variables X, Y, and Z^S.

Imposing decomposition constraint

- Maximizing interaction information to encode shared information

encoding information shared between X and Y to Z^S

maximize $$I(X;Y;Z^S) = I\left(X;Z^S\right) - I\left(X;Z^S|Y\right)$$

$$= I\left(Y;Z^S\right) - I\left(Y;Z^S|X\right)$$ (due to symmetry)

Imposing decomposition constraint

- Maximizing interaction information to encode shared information

encoding information shared between X and Y to Z^S

maximize $I(X;Y;Z^S) = I\left(X;Z^S\right) - I\left(X;Z^S|Y\right)$

Z^S should be informative to X

Z^S should be also inferable from Y

Z^S should encode maximum information shared between X and Y

Joint regularization

$$\max_q \mathcal{L}_{disentangle}(q) = 2 \cdot \underbrace{I(X;Y;Z^S)}_{} - I(Z^X;Z^S) - I(Z^Y;Z^S)$$

Interaction information:

The amount of information shared among three variables X, Y, and Z^S.

The maximization encourages Z^S to capture only the shared factors of variation.

=> Decomposition

Joint regularization

$$\max_q \mathcal{L}_{disentangle}(q) = 2 \cdot I(X;Y;Z^S) - \underbrace{I(Z^X;Z^S)}_{} - I(Z^Y;Z^S)$$

Mutual information:

The amount of information shared between two variables Z^X and Z^S.

The minimization makes Z^X and Z^S independent.

Joint regularization

$$\max_{q} \mathcal{L}_{disentangle}(q) = 2 \cdot I(X;Y;Z^S) - I(Z^X;Z^S) - \underbrace{I(Z^Y;Z^S)}_{\textbf{Mutual information:}}$$

The amount of information shared between two variables Z^Y and Z^S.

The minimization makes Z^Y and Z^S independent.

=> Disentanglement

Comparing lower-bounds

- Lower-bound of VAE objective

$$\mathcal{L}_{ELBO}(p,q)$$
$$\geq \mathbb{E}_{q(z^x|x)q(z^s|x,y)}\left[\log p(x|z^x,z^s)\right]$$
$$+\mathbb{E}_{q(z^y|y)q(z^s|x,y)}\left[\log p(y|z^y,z^s)\right]$$
$$-D_{KL}\left[q(z^x|x)\|p(z^x)\right]$$
$$-D_{KL}\left[q(z^y|y)\|p(z^y)\right]$$
$$-D_{KL}\left[q(z^s|x,y)\|p(z^s)\right]$$

- Lower-bound of regularization

$$\mathcal{L}_{\text{disentangle}}(p,q,r)$$
$$\geq \mathbb{E}_{q(z^s|x,y)q(z^x|x)}\left[\log p(x|z^x,z^s)\right]$$
$$+\mathbb{E}_{q(z^s|x,y)q(z^y|y)}\left[\log p(y|z^y,z^s)\right]$$
$$-D_{KL}\left[q(z^x|x)\|p(z^x)\right]$$
$$-D_{KL}\left[q(z^y|y)\|p(z^y)\right]$$
$$-D_{KL}\left[q(z^s|x,y)\|r^y(z^s|y)\right]$$
$$-D_{KL}\left[q(z^s|x,y)\|r^x(z^s|x)\right]$$

Surprisingly, same terms appear in both objectives

Interaction Information AutoEncoder (IIAE)

- Objective function

$$
\begin{aligned}
&\max_{p,q} \mathcal{L}_{ELBO}(p,q) + \lambda \cdot \mathcal{L}_{disentangle}(q) \\
&\geq \max_{p,q,r} (1+\lambda) \cdot \ ELBO(p,q) \\
&\qquad\qquad + \lambda \cdot \ D_{KL}\left[q(z^s|x,y) \| p(z^s)\right] \\
&\qquad\qquad - \lambda \cdot (\ D_{KL}\left[q(z^s|x,y) \| r^y(z^s|y)\right] + D_{KL}\left[q(z^s|x,y) \| r^x(z^s|x)\right]\ ) \,.
\end{aligned}
$$

- Advantages

Interaction Information AutoEncoder (IIAE)

- Objective function

$$\max_{p,q} \mathcal{L}_{ELBO}(p,q) + \lambda \cdot \mathcal{L}_{disentangle}(q)$$

$$\geq \max_{p,q,r} (1+\lambda) \cdot \ ELBO(p,q)$$

$$+ \lambda \cdot \ D_{KL}\left[q(z^s|x,y) \| p(z^s)\right]$$

$$- \lambda \cdot \left(D_{KL}\left[q(z^s|x,y) \| r^y(z^s|y)\right] + D_{KL}\left[q(z^s|x,y) \| r^x(z^s|x)\right] \right).$$

- Advantages
  - IIAE Introduces only two additional terms for regularization

Interaction Information AutoEncoder (IIAE)

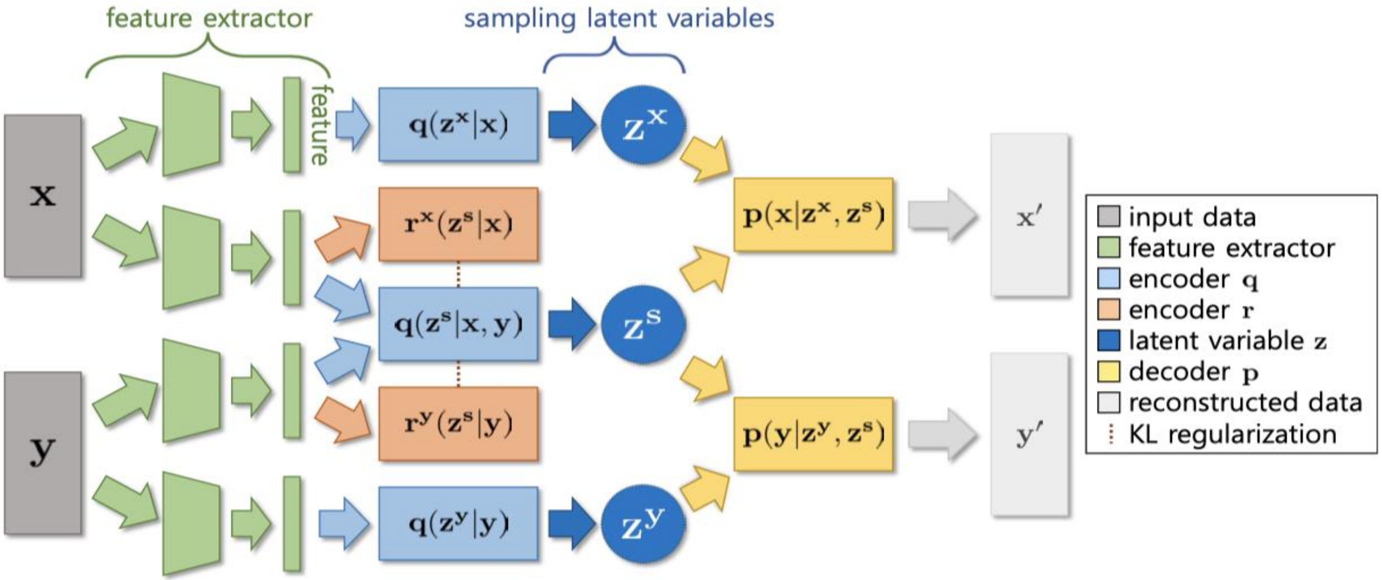
- Objective function

$$
\begin{aligned}
&\max_{p,q} \mathcal{L}_{ELBO}(p,q) + \lambda \cdot \mathcal{L}_{disentangle}(q) \\
&\geq \max_{p,q,r} (1+\lambda) \cdot \ ELBO(p,q) \\
&\qquad\qquad + \lambda \cdot \ D_{KL}\left[q(z^s|x,y) \| p(z^s)\right] \\
&\qquad\qquad - \lambda \cdot \left(\ D_{KL}\left[q(z^s|x,y) \| r^y(z^s|y)\right] + D_{KL}\left[q(z^s|x,y) \| r^x(z^s|x)\right] \ \right).
\end{aligned}
$$

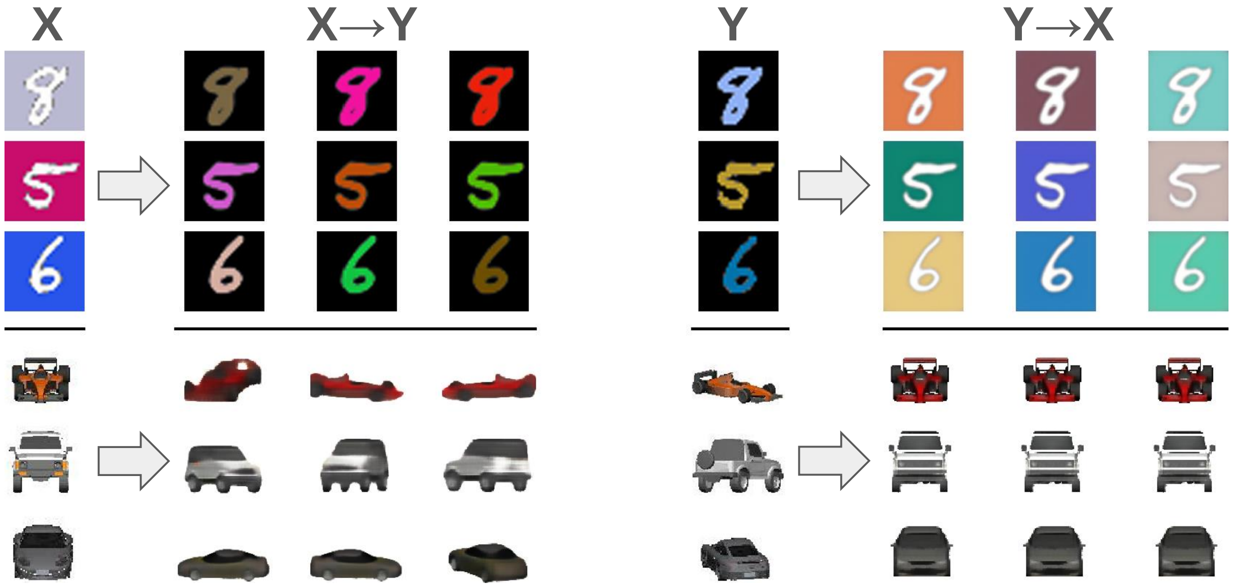
- Advantages
  - IIAE Introduces only two additional terms for regularization
  - Shared representation can be extracted by either x or y (it does not require both)

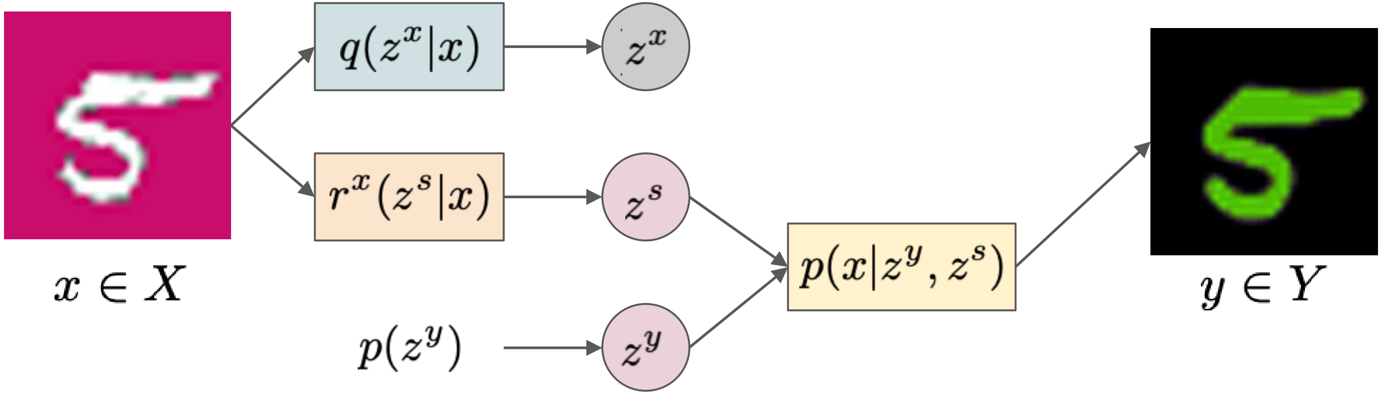
Interaction Information AutoEncoder (IIAE)

- Overall architecture:

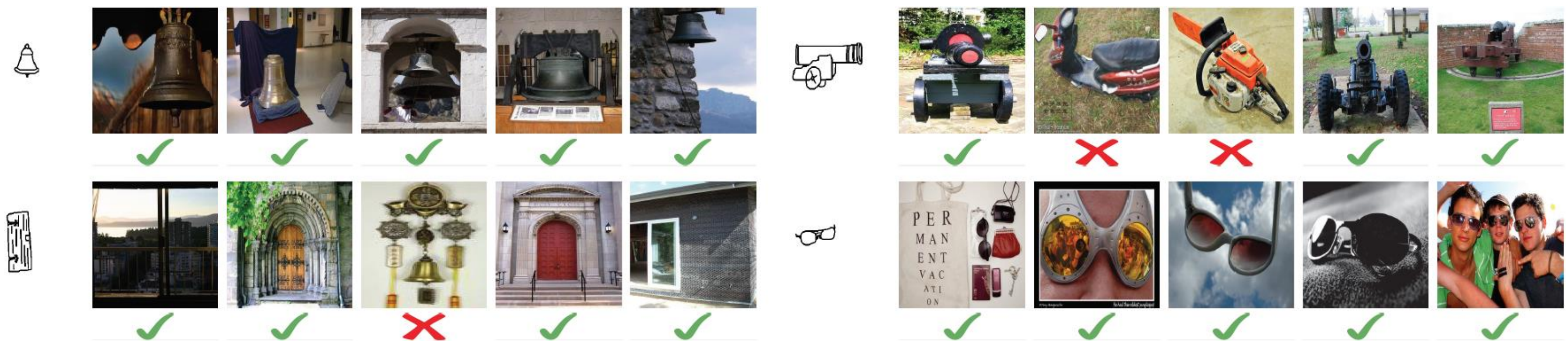


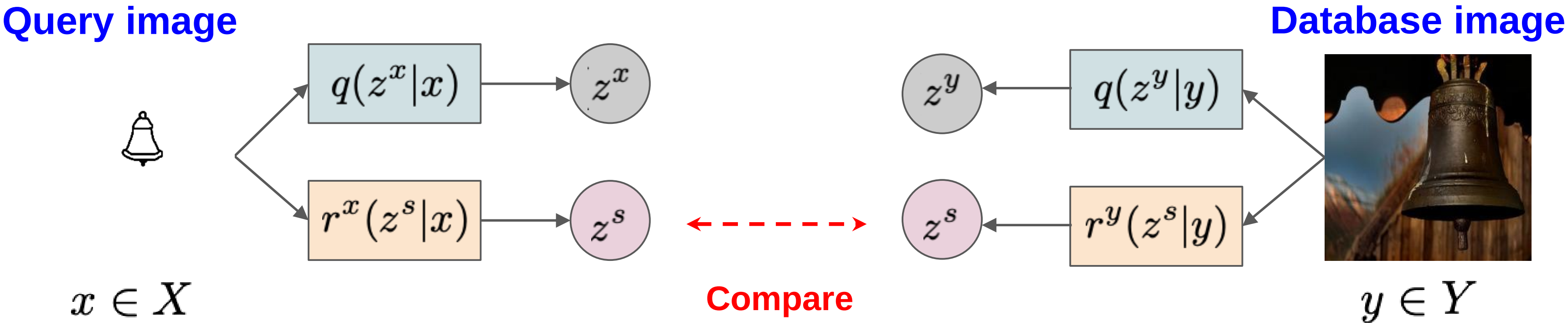
Task 1: Cross-domain Image Translation





Task 2: Zero Shot – Sketch Based Image Retrieval





Task 2: Zero Shot – Sketch Based Image Retrieval

Models	Feature Dimension	Evaluation metric		External knowledge		
		mAP	P@100	Attr.	WordEmb.	WordNet [33]
SAE [23]	300	0.216	0.293	✓	✓	-
FRWGAN [9]	512	0.127	0.169	✓	-	-
ZSIH [38]	64	0.258	0.342	-	✓	-
CAAE [22]	4096	0.196	0.284	-	-	-
SEM-PCYC [6]	64	0.349	0.463	-	✓	✓
LCALE [27]	64	0.476	0.583	-	✓	-
IIAE	64	**0.573**	**0.659**	-	-	-

Table 3: Evaluation on the Sketchy Extended dataset [29, 37]. Attr and WordEmb stand for attribute information and word embedding respectively.