Generative Modeling with Succinct Representation Learning via Wyner's Common Information

Jongha (Jon) Ryu
MIT

January 18, 2023

Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations

Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{(\text{story}_i, \text{illustration}_i)\}$,

Alice was beginning to get very tired of sitting ...when suddenly a White Rabbit with pink eyes ran close by her ...see it pop down a large rabbit-hole under the hedge.

Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{$(story$_i$, illustration$_i$)$\}$, $\{$(image$_i$, caption$_i$)$\}$, ...

the bird has a white body, black wings, and webbed orange feet

a blue bird with gray primaries and secondaries and white breast and throat

Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)

Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
  - Joint generation: Learn $q(\mathbf{x}, \mathbf{y})$ and generate $(\mathbf{X}, \mathbf{Y})$

Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
  - Joint generation: Learn $q(\mathbf{x}, \mathbf{y})$ and generate $(\mathbf{X}, \mathbf{Y})$
  - Conditional generation: Learn $q(\mathbf{y}|\mathbf{x})$ and generate $\mathbf{Y}$ given $\mathbf{x} \sim q(\mathbf{x})$

Problem setting

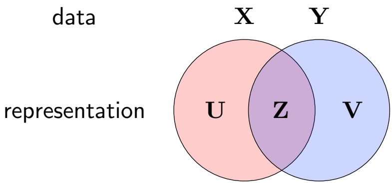
- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
  - Joint generation: Learn $q(\mathbf{x}, \mathbf{y})$ and generate $(\mathbf{X}, \mathbf{Y})$
  - Conditional generation: Learn $q(\mathbf{y}|\mathbf{x})$ and generate $\mathbf{Y}$ given $\mathbf{x} \sim q(\mathbf{x})$
  - Cross-domain retrieval: Given a query $\mathbf{x}$, retrieve relevant $\mathbf{y}$'s from a pool $\{\mathbf{y}_i\}_{i=1}^n$

Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
- Fit a generative model with disentangled latent representations $(\mathbf{Z}, \mathbf{U}, \mathbf{V})$

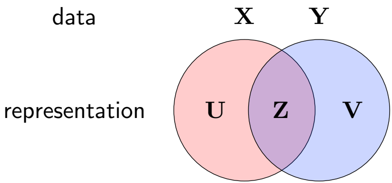
Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{(\text{story}_i, \text{illustration}_i)\}$, $\{(\text{image}_i, \text{caption}_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
- Fit a generative model with disentangled latent representations $(\mathbf{Z}, \mathbf{U}, \mathbf{V})$
  - Disentangle commonality $\mathbf{Z}$ from private properties $\mathbf{U}$ and $\mathbf{V}$



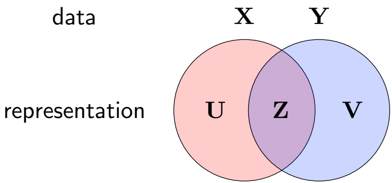
Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
- Fit a generative model with disentangled latent representations $(\mathbf{Z}, \mathbf{U}, \mathbf{V})$
  - Disentangle commonality $\mathbf{Z}$ from private properties $\mathbf{U}$ and $\mathbf{V}$
  - a.k.a. cross-domain disentanglement problem



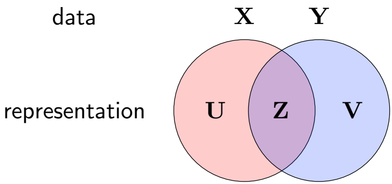
Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
- Fit a generative model with disentangled latent representations $(\mathbf{Z}, \mathbf{U}, \mathbf{V})$
- **Q**. Under which criterion should we disentangle? What is an optimal representation?



Problem setting

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i)\}$ i.i.d. $\sim q(\mathbf{x}, \mathbf{y})$; high. dim., many-to-many relations
- **Examples**: $\{($story$_i$, illustration$_i)\}$, $\{($image$_i$, caption$_i)\}$, ...
- **Goal**: learn the data distribution and sample from it (a.k.a. generation)
- Fit a generative model with disentangled latent representations $(\mathbf{Z}, \mathbf{U}, \mathbf{V})$
- **Q**. Under which criterion should we disentangle? What is an optimal representation?
- **Proposal**: use Wyner's common information to learn succinct common representation



Motivation

- Cooperative game between Alice and Bob

Motivation

- Cooperative game between Alice and Bob

Motivation

- Cooperative game between Alice and Bob

Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo

$\mathbf{X}$

Child's photo

Motivation

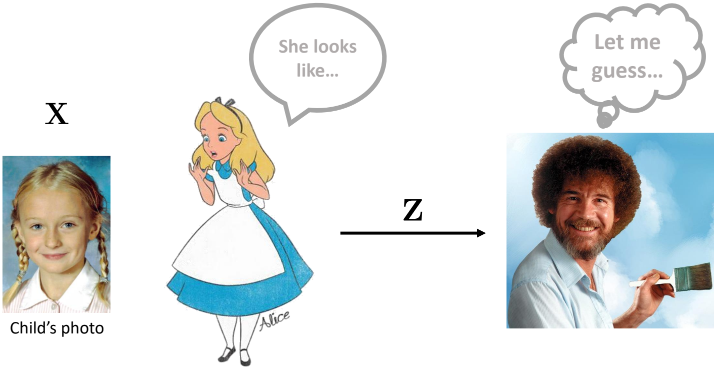
- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo

$\mathbf{X}$

Child's photo

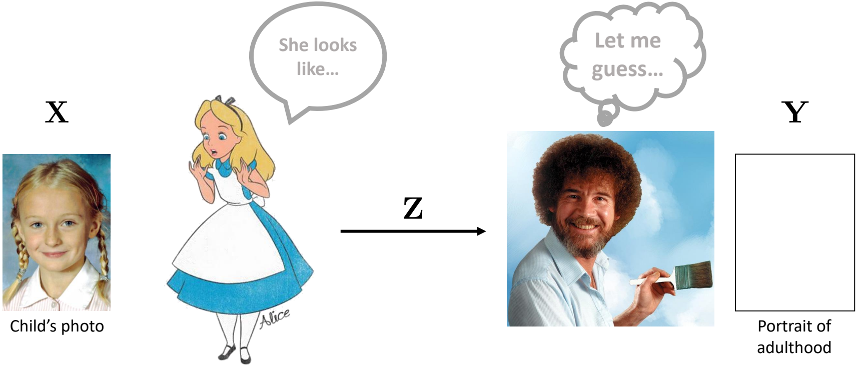
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo



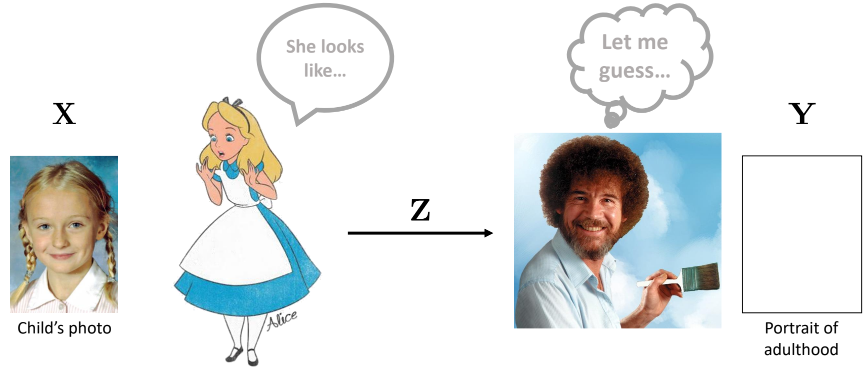
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo



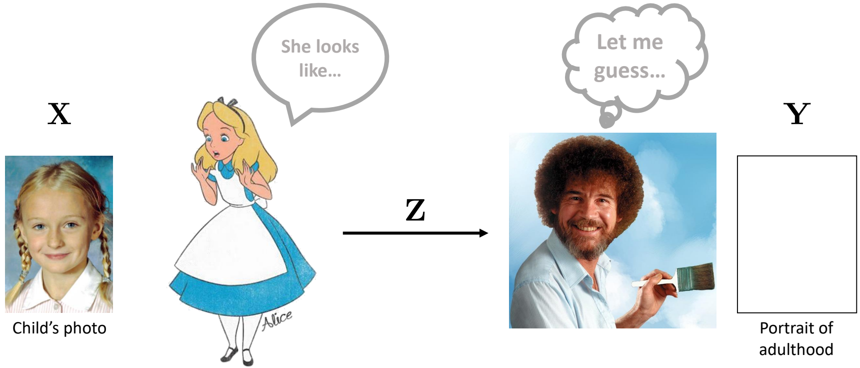
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo



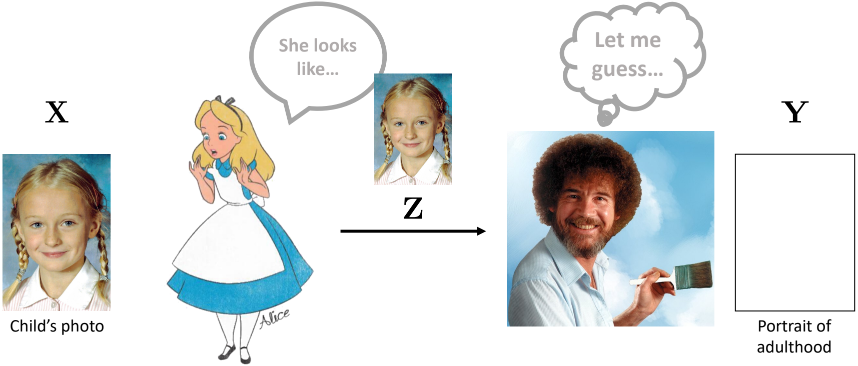
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?



Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?

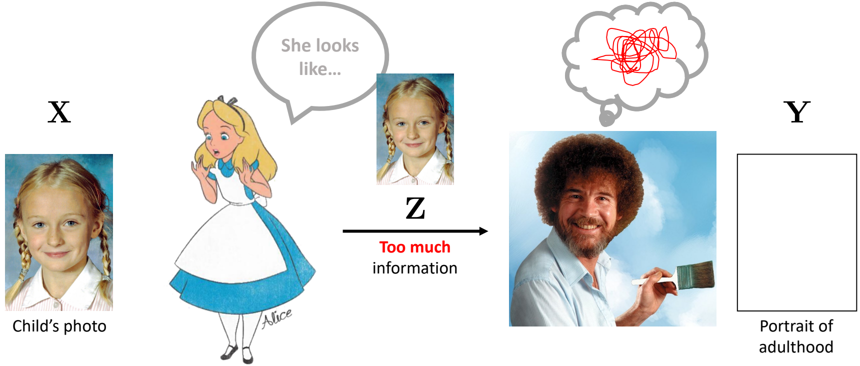


Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?

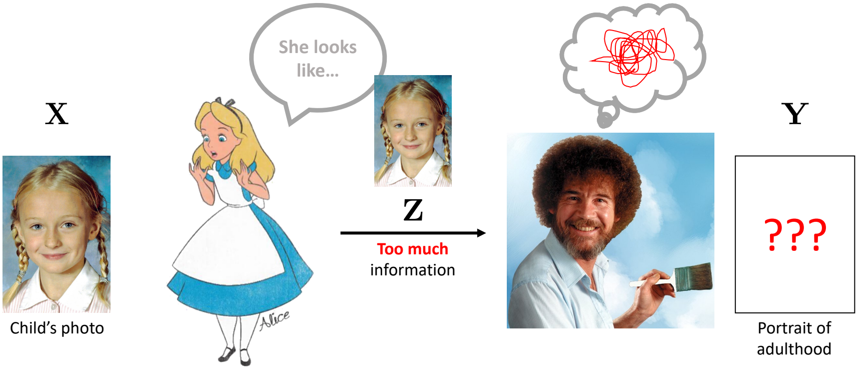
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?



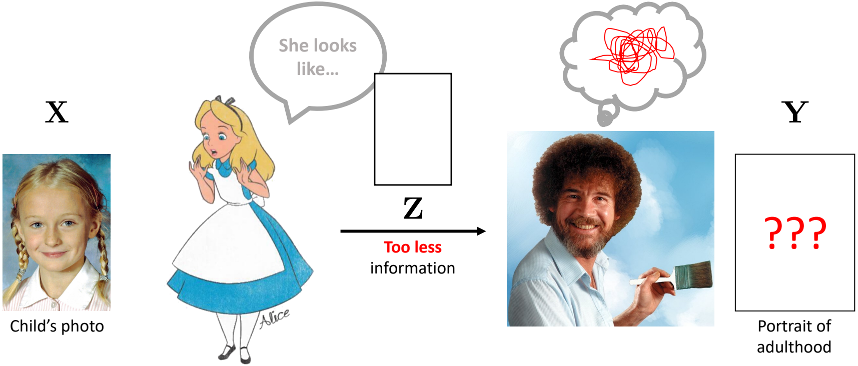
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?



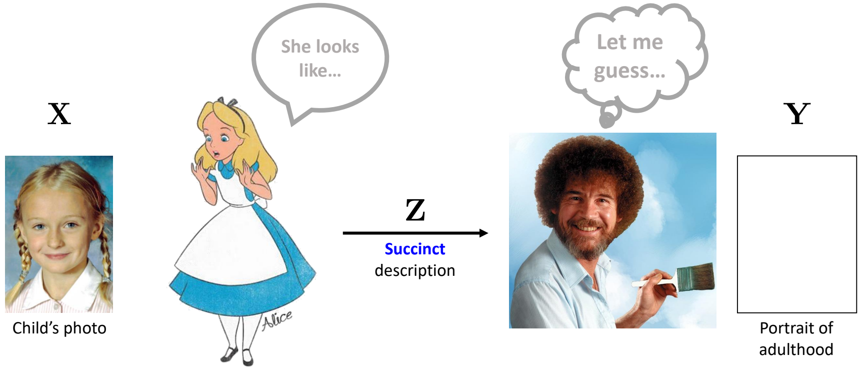
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?



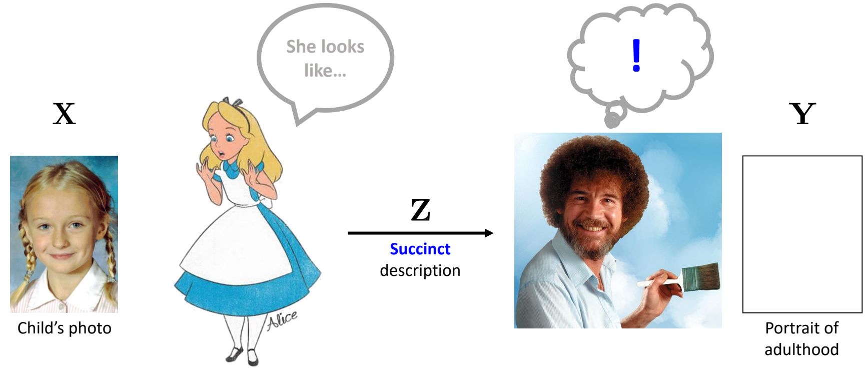
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?



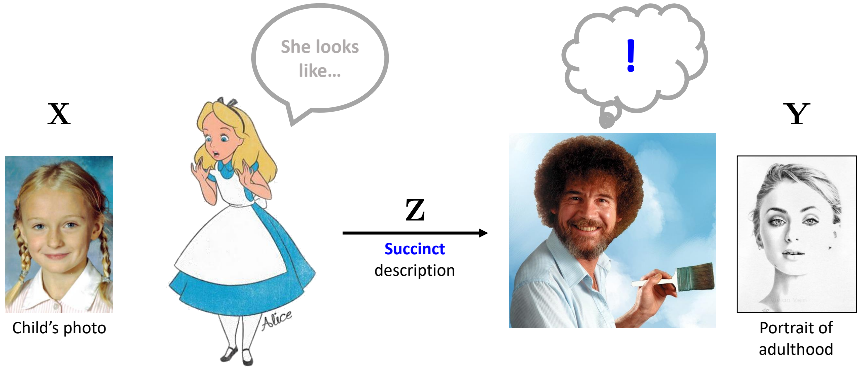
Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?

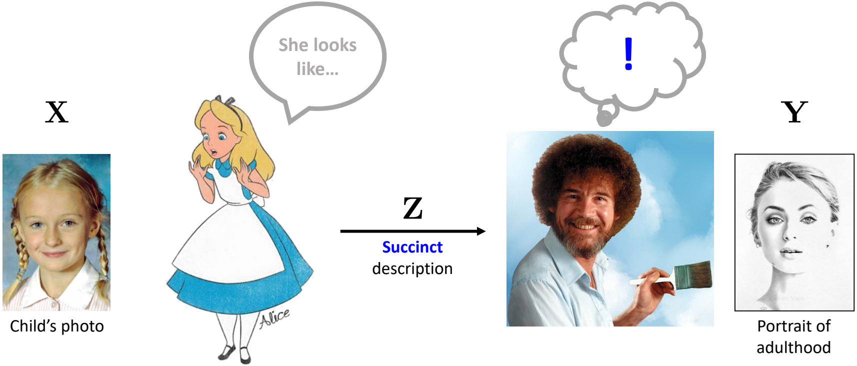


Motivation

- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?

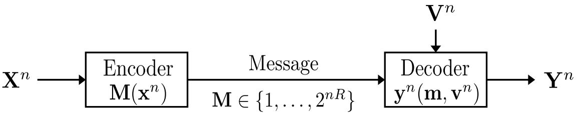


- Cooperative game between Alice and Bob
- Alice and Bob wish to draw a nice portrait of adulthood from a child's photo
- What description dose Alice need to generate and send?
- Alice can maximally help Bob by providing the most "succinct" description!



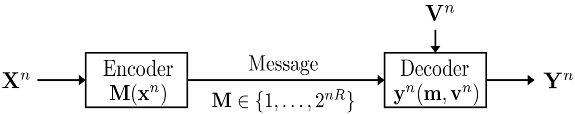
Channel synthesis (Cuff 2013)

- **Problem**: simulate a channel $q(\mathbf{y}|\mathbf{x})$ by communicating nR bits



Channel synthesis (Cuff 2013)

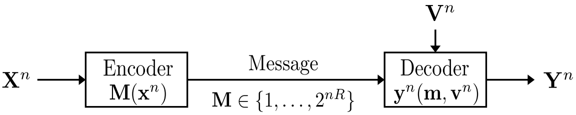
- **Problem**: simulate a channel $q(\mathbf{y}|\mathbf{x})$ by communicating nR bits



- **Question**: What is the minimum rate R^*?

Channel synthesis (Cuff 2013)

- **Problem**: simulate a channel $q(\mathbf{y}|\mathbf{x})$ by communicating nR bits

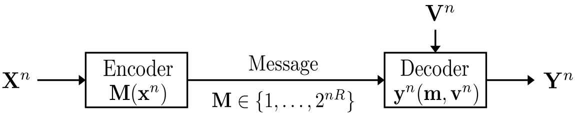


- **Question**: What is the minimum rate R^*?
- **Answer**: Wyner's common information $R^* = J(\mathbf{X}; \mathbf{Y})$

minimize	$I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q(\mathbf{z}\|\mathbf{x}, \mathbf{y})$

Channel synthesis (Cuff 2013)

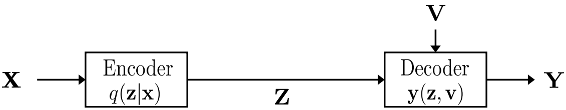
- **Problem**: simulate a channel $q(\mathbf{y}|\mathbf{x})$ by communicating nR bits



- **Question**: What is the minimum rate R^*?
- **Answer**: Wyner's common information $R^* = J(\mathbf{X}; \mathbf{Y})$

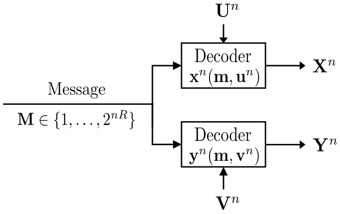
minimize	$I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q(\mathbf{z}\|\mathbf{x}, \mathbf{y})$

- Single-letter characterization



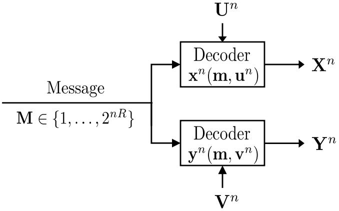
Distributed simulation (Wyner 1975)

- **Problem**: simulate a joint distribution $q(\mathbf{x}, \mathbf{y})$ from nR common bits



Distributed simulation (Wyner 1975)

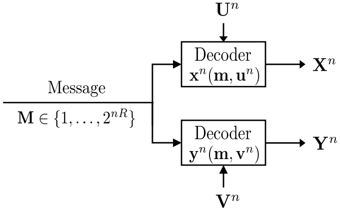
- **Problem**: simulate a joint distribution $q(\mathbf{x}, \mathbf{y})$ from nR common bits



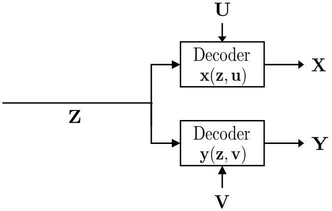
- **Question**: What is the minimum rate R^{**}?

Distributed simulation (Wyner 1975)

- **Problem**: simulate a joint distribution $q(\mathbf{x}, \mathbf{y})$ from nR common bits



- **Question**: What is the minimum rate R^{**}?
- **Answer**: $R^{**} = J(\mathbf{X}; \mathbf{Y})$



Learning distributions based on Wyner's common information

- Channel synthesis $\rightarrow$ conditional generation
- Distributed simulation $\rightarrow$ joint generation
- Joint and conditional distributions share the same common information structure

minimize	$I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$

Learning distributions based on Wyner's common information

- Channel synthesis $\rightarrow$ conditional generation
- Distributed simulation $\rightarrow$ joint generation
- Joint and conditional distributions share the same common information structure

minimize	$I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q_\phi(\mathbf{z}\vert\mathbf{x}, \mathbf{y})$

- **Proposal**: use the optimization problem as a principle of CD disentanglement!

Learning distributions based on Wyner's common information

- Channel synthesis $\rightarrow$ conditional generation
- Distributed simulation $\rightarrow$ joint generation
- Joint and conditional distributions share the same common information structure

minimize	$I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$

- **Proposal**: use the optimization problem as a principle of CD disentanglement!
- $I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$ quantifies the complexity of the representation $\mathbf{Z}$

Learning distributions based on Wyner's common information

- Channel synthesis $\rightarrow$ conditional generation
- Distributed simulation $\rightarrow$ joint generation
- Joint and conditional distributions share the same common information structure

$$\begin{array}{ll} \text{minimize} & I(\mathbf{X}, \mathbf{Y}; \mathbf{Z}) \\ \text{subject to} & \mathbf{X} - \mathbf{Z} - \mathbf{Y} \\ \text{variables} & q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}) \end{array}$$

- **Proposal**: use the optimization problem as a principle of CD disentanglement!
- $I(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$ quantifies the complexity of the representation $\mathbf{Z}$
- **Goal**: learn a generative model with Wyner's common information structure

Learning distributions based on Wyner's common information

- Channel synthesis $\to$ conditional generation
- Distributed simulation $\to$ joint generation
- Joint and conditional distributions share the same common information structure

minimize	$I(\mathbf{X},\mathbf{Y};\mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q_\phi(\mathbf{z}\|\mathbf{x},\mathbf{y})$

- **Proposal**: use the optimization problem as a principle of CD disentanglement!
- $I(\mathbf{X},\mathbf{Y};\mathbf{Z})$ quantifies the complexity of the representation $\mathbf{Z}$
- **Goal**: learn a generative model with Wyner's common information structure
- We consider the latent variable models induced by the single letter characterizations

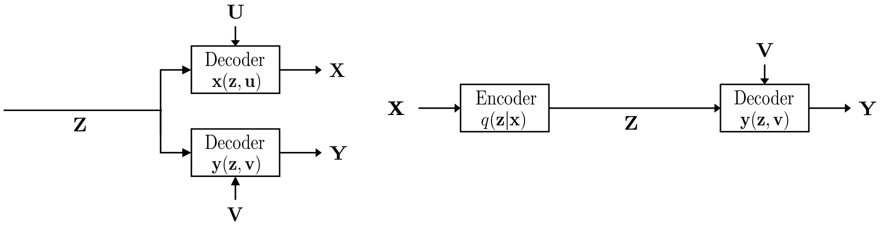
Learning distributions based on Wyner's common information

- Channel synthesis → conditional generation
- Distributed simulation → joint generation
- Joint and conditional distributions share the same common information structure

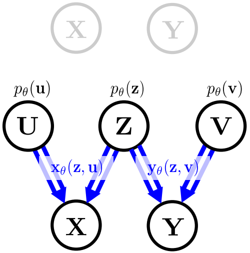
minimize	$I(\mathbf{X},\mathbf{Y};\mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q_\phi(\mathbf{z}\|\mathbf{x},\mathbf{y})$

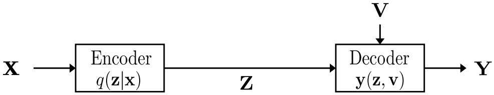
- **Proposal**: use the optimization problem as a principle of CD disentanglement!
- $I(\mathbf{X},\mathbf{Y};\mathbf{Z})$ quantifies the complexity of the representation $\mathbf{Z}$
- **Goal**: learn a generative model with Wyner's common information structure
- We consider the latent variable models induced by the single letter characterizations
- We will fit the generative models to data based on Wyner's optimization problem

Probabilistic model for joint and conditional sampling

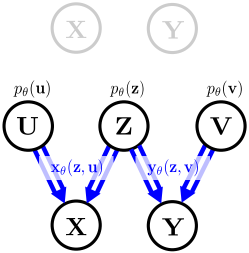


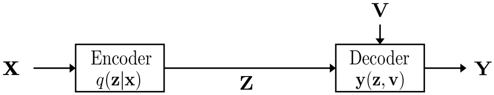
Probabilistic model for joint and conditional sampling





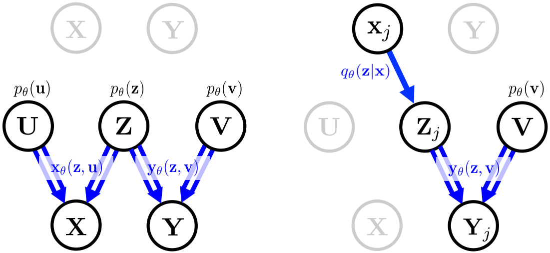
Probabilistic model for joint and conditional sampling





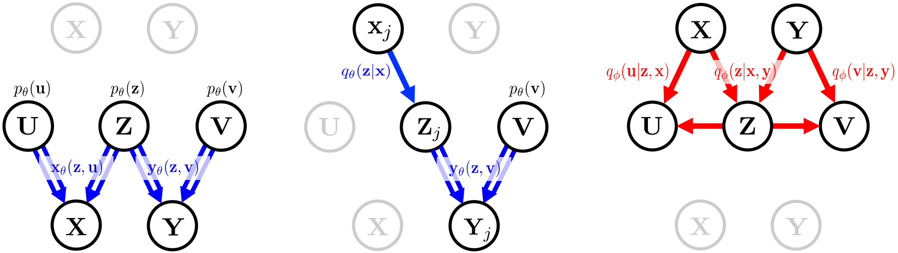
- Decoders: $\mathbf{x}_\theta(\mathbf{z},\mathbf{u}), \mathbf{y}_\theta(\mathbf{z},\mathbf{v})$
- Priors (source of randomness): common $p_\theta(\mathbf{z})$, local $p_\theta(\mathbf{u}), p_\theta(\mathbf{v})$

Probabilistic model for joint and conditional sampling



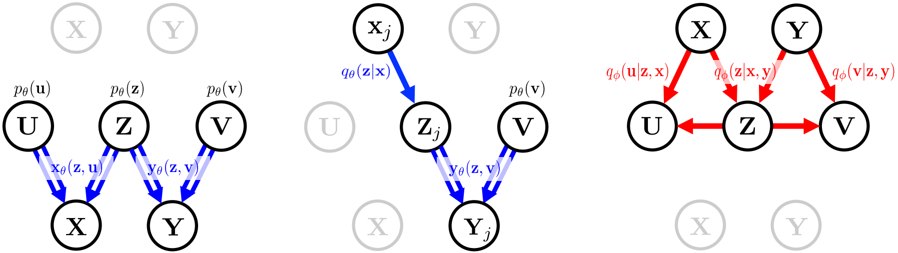
- Decoders: $\mathbf{x}_\theta(\mathbf{z},\mathbf{u}), \mathbf{y}_\theta(\mathbf{z},\mathbf{v})$
- Priors (source of randomness): common $p_\theta(\mathbf{z})$, local $p_\theta(\mathbf{u}), p_\theta(\mathbf{v})$
- Model (marginal) encoders: $q_\theta(\mathbf{z}|\mathbf{x})$, $q_\theta(\mathbf{z}|\mathbf{y})$

Probabilistic model for joint and conditional sampling



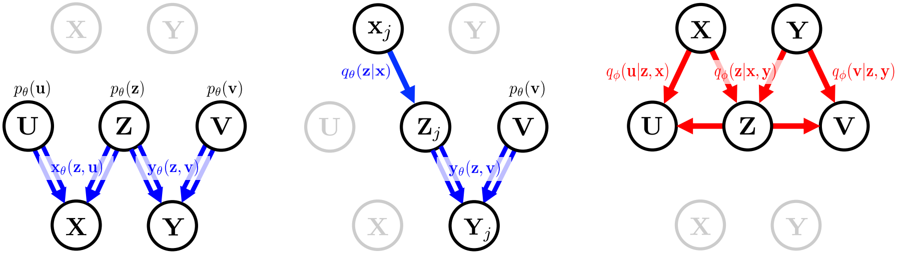
- Decoders: $\mathbf{x}_\theta(\mathbf{z},\mathbf{u}), \mathbf{y}_\theta(\mathbf{z},\mathbf{v})$
- Priors (source of randomness): common $p_\theta(\mathbf{z})$, local $p_\theta(\mathbf{u}), p_\theta(\mathbf{v})$
- Model (marginal) encoders: $q_\theta(\mathbf{z}|\mathbf{x})$, $q_\theta(\mathbf{z}|\mathbf{y})$
- Variational encoders: joint $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, local $q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})$, $q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$

Probabilistic model for joint and conditional sampling



- Decoders: $\mathbf{x}_\theta(\mathbf{z},\mathbf{u}), \mathbf{y}_\theta(\mathbf{z},\mathbf{v})$
- Priors (source of randomness): common $p_\theta(\mathbf{z})$, local $p_\theta(\mathbf{u}), p_\theta(\mathbf{v})$
- Model (marginal) encoders: $q_\theta(\mathbf{z}|\mathbf{x})$, $q_\theta(\mathbf{z}|\mathbf{y})$
- Variational encoders: joint $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, local $q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})$, $q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$
- Call these components in entirety the variational Wyner model

Probabilistic model for joint and conditional sampling



- Decoders: $\mathbf{x}_\theta(\mathbf{z},\mathbf{u}), \mathbf{y}_\theta(\mathbf{z},\mathbf{v})$
- Priors (source of randomness): common $p_\theta(\mathbf{z})$, local $p_\theta(\mathbf{u}), p_\theta(\mathbf{v})$
- Model (marginal) encoders: $q_\theta(\mathbf{z}|\mathbf{x})$, $q_\theta(\mathbf{z}|\mathbf{y})$

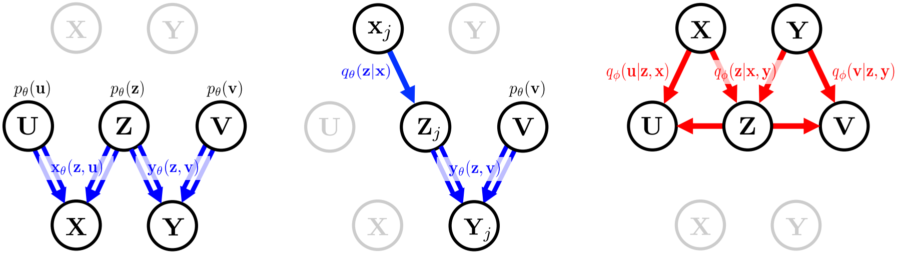
} model θ

- Variational encoders: joint $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, local $q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})$, $q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$

} variational ϕ

- Call these components in entirety the variational Wyner model

Probabilistic model for joint and conditional sampling



- Decoders: $\mathbf{x}_\theta(\mathbf{z},\mathbf{u}), \mathbf{y}_\theta(\mathbf{z},\mathbf{v})$
- Priors (source of randomness): common $p_\theta(\mathbf{z})$, local $p_\theta(\mathbf{u}), p_\theta(\mathbf{v})$

} decoders p

- Model (marginal) encoders: $q_\theta(\mathbf{z}|\mathbf{x})$, $q_\theta(\mathbf{z}|\mathbf{y})$
- Variational encoders: joint $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, local $q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x})$, $q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$

} encoders q

- Call these components in entirety the variational Wyner model

Training objectives

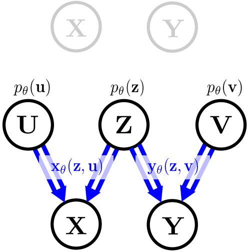
- The variational Wyner model induces four distributions:

joint
cond. $(x \to y)$
cond. $(y \to x)$
variational

Training objectives

- The variational Wyner model induces four distributions:

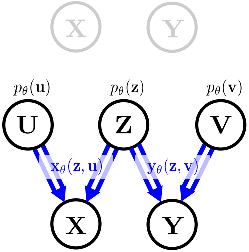
joint	$p_{\to \mathsf{xy}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (x → y)	
cond. (y → x)	
variational	

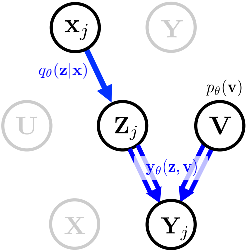


Training objectives

- The variational Wyner model induces four distributions:

joint	$p_{\to \mathsf{xy}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (x $\to$ y)	$p_{\mathsf{x}\to\mathsf{y}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}) \triangleq q(\mathbf{x})q_\theta(\mathbf{z}\|\mathbf{x})p_\theta(\mathbf{v})\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (y $\to$ x)	
variational	

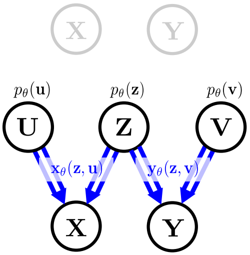


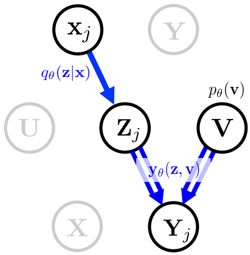


Training objectives

- The variational Wyner model induces four distributions:

joint	$p_{\to \mathsf{xy}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (x $\to$ y)	$p_{\mathsf{x}\to\mathsf{y}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}) \triangleq q(\mathbf{x})q_\theta(\mathbf{z}\|\mathbf{x})p_\theta(\mathbf{v})\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (y $\to$ x)	$p_{\mathsf{y}\to\mathsf{x}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u}) \triangleq q(\mathbf{y})q_\theta(\mathbf{z}\|\mathbf{y})p_\theta(\mathbf{u})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))$
variational	

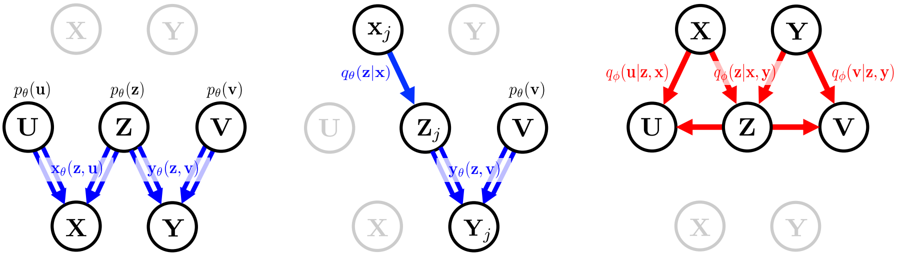




Training objectives

- The variational Wyner model induces four distributions:

joint	$p_{\to \mathsf{xy}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (x → y)	$p_{\mathsf{x}\to\mathsf{y}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}) \triangleq q(\mathbf{x})q_\theta(\mathbf{z}\mid\mathbf{x})p_\theta(\mathbf{v})\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (y → x)	$p_{\mathsf{y}\to\mathsf{x}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u}) \triangleq q(\mathbf{y})q_\theta(\mathbf{z}\mid\mathbf{y})p_\theta(\mathbf{u})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))$
variational	$q_{\mathsf{xy}\to}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}\mid\mathbf{x},\mathbf{y})q_\phi(\mathbf{u}\mid\mathbf{z},\mathbf{x})q_\phi(\mathbf{v}\mid\mathbf{z},\mathbf{y})$



Training objectives

- The variational Wyner model induces four distributions:

joint	$p_{\to \mathsf{xy}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. $(\mathsf{x}\to\mathsf{y})$	$p_{\mathsf{x}\to\mathsf{y}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}) \triangleq q(\mathbf{x})q_\theta(\mathbf{z}\|\mathbf{x})p_\theta(\mathbf{v})\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. $(\mathsf{y}\to\mathsf{x})$	$p_{\mathsf{y}\to\mathsf{x}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u}) \triangleq q(\mathbf{y})q_\theta(\mathbf{z}\|\mathbf{y})p_\theta(\mathbf{u})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))$
variational	$q_{\mathsf{xy}\to}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}\|\mathbf{x},\mathbf{y})q_\phi(\mathbf{u}\|\mathbf{z},\mathbf{x})q_\phi(\mathbf{v}\|\mathbf{z},\mathbf{y})$

- Recall Wyner's optimization problem:

minimize	$I(\mathbf{X},\mathbf{Y};\mathbf{Z})$
subject to	$\mathbf{X}-\mathbf{Z}-\mathbf{Y}$
variables	$q_\phi(\mathbf{z}\|\mathbf{x},\mathbf{y})$

Training objectives

- The variational Wyner model induces four distributions:

joint	$p_{\to \mathsf{xy}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v}) \triangleq p_\theta(\mathbf{z}) p_\theta(\mathbf{u}) p_\theta(\mathbf{v}) \delta(\mathbf{x} - \mathbf{x}_\theta(\mathbf{z}, \mathbf{u})) \delta(\mathbf{y} - \mathbf{y}_\theta(\mathbf{z}, \mathbf{v}))$
cond. $(\mathsf{x} \to \mathsf{y})$	$p_{\mathsf{x} \to \mathsf{y}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{v}) \triangleq q(\mathbf{x}) q_\theta(\mathbf{z}\|\mathbf{x}) p_\theta(\mathbf{v}) \delta(\mathbf{y} - \mathbf{y}_\theta(\mathbf{z}, \mathbf{v}))$
cond. $(\mathsf{y} \to \mathsf{x})$	$p_{\mathsf{y} \to \mathsf{x}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}) \triangleq q(\mathbf{y}) q_\theta(\mathbf{z}\|\mathbf{y}) p_\theta(\mathbf{u}) \delta(\mathbf{x} - \mathbf{x}_\theta(\mathbf{z}, \mathbf{u}))$
variational	$q_{\mathsf{xy} \to}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v}) \triangleq q(\mathbf{x}, \mathbf{y}) q_\phi(\mathbf{z}\|\mathbf{x}, \mathbf{y}) q_\phi(\mathbf{u}\|\mathbf{z}, \mathbf{x}) q_\phi(\mathbf{v}\|\mathbf{z}, \mathbf{y})$

- Recall Wyner's optimization problem:

$$\begin{array}{ll} \text{minimize} & I(\mathbf{X}, \mathbf{Y}; \mathbf{Z}) \\ \text{subject to} & \mathbf{X} - \mathbf{Z} - \mathbf{Y} \\ \text{variables} & q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}) \end{array}$$

- For each model $p_{\mathsf{model}} \in \{p_{\to \mathsf{xy}}, p_{\mathsf{x} \to \mathsf{y}}, p_{\mathsf{y} \to \mathsf{x}}\}$, we can relax the problem as

$$\text{minimize} \quad D(p_{\mathsf{model}}, q_{\mathsf{xy} \to}) + \lambda_{\mathsf{model}}^{\mathsf{CI}} I_{\mathsf{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$$

Training objectives

- The variational Wyner model induces four distributions:

joint	$p_{\to xy}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (x $\to$ y)	$p_{x\to y}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}) \triangleq q(\mathbf{x})q_\theta(\mathbf{z}\|\mathbf{x})p_\theta(\mathbf{v})\delta(\mathbf{y}-\mathbf{y}_\theta(\mathbf{z},\mathbf{v}))$
cond. (y $\to$ x)	$p_{y\to x}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u}) \triangleq q(\mathbf{y})q_\theta(\mathbf{z}\|\mathbf{y})p_\theta(\mathbf{u})\delta(\mathbf{x}-\mathbf{x}_\theta(\mathbf{z},\mathbf{u}))$
variational	$q_{xy\to}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}) \triangleq q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}\|\mathbf{x},\mathbf{y})q_\phi(\mathbf{u}\|\mathbf{z},\mathbf{x})q_\phi(\mathbf{v}\|\mathbf{z},\mathbf{y})$

- Recall Wyner's optimization problem:

$$\begin{array}{ll} \text{minimize} & I(\mathbf{X},\mathbf{Y};\mathbf{Z}) \\ \text{subject to} & \mathbf{X}-\mathbf{Z}-\mathbf{Y} \\ \text{variables} & q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) \end{array}$$

- For each model $p_{\text{model}} \in \{p_{\to xy}, p_{x\to y}, p_{y\to x}\}$, we can relax the problem as

$$\text{minimize} \quad D(p_{\text{model}}, q_{xy\to}) + \lambda^{\text{CI}}_{\text{model}} I_{\text{model}}(\mathbf{X},\mathbf{Y};\mathbf{Z})$$

- Distribution matching with CI regularization

Derivation

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$:

minimize	$I_{\text{xy}\to}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$

Derivation

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$:

minimize	$I_{\text{xy}\to}(\mathbf{X},\mathbf{Y};\mathbf{Z})$
subject to	$\mathbf{X} - \mathbf{Z} - \mathbf{Y}$
variables	$q_\phi(\mathbf{z}\|\mathbf{x},\mathbf{y})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency

Derivation

- For each model $p_{\text{model}} \in \{p_{\rightarrow\text{xy}}, p_{\text{x}\rightarrow\text{y}}, p_{\text{y}\rightarrow\text{x}}\}$:

minimize	$I_{\text{xy}\rightarrow}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$p_{\text{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v}) \equiv q_{\text{xy}\rightarrow}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})$
variables	$q_\phi(\mathbf{z}\|\mathbf{x}, \mathbf{y}), q_\phi(\mathbf{u}\|\mathbf{z}, \mathbf{x}), q_\phi(\mathbf{v}\|\mathbf{z}, \mathbf{y}), p_{\text{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency

Derivation

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$:

minimize	$I_{\text{xy}\to}(\mathbf{X},\mathbf{Y};\mathbf{Z})$
subject to	$D(p_{\text{model}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}), q_{\text{xy}\to}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})) = 0$
variables	$q_\phi(\mathbf{z}\vert\mathbf{x},\mathbf{y}), q_\phi(\mathbf{u}\vert\mathbf{z},\mathbf{x}), q_\phi(\mathbf{v}\vert\mathbf{z},\mathbf{y}), p_{\text{model}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency

Derivation

- For each model $p_{\mathsf{model}} \in \{p_{\rightarrow\mathsf{xy}}, p_{\mathsf{x}\rightarrow\mathsf{y}}, p_{\mathsf{y}\rightarrow\mathsf{x}}\}$:

minimize	$I_{\mathsf{xy}\rightarrow}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$D(p_{\mathsf{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v}), q_{\mathsf{xy}\rightarrow}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})) = 0$
variables	$q_\phi(\mathbf{z}\|\mathbf{x}, \mathbf{y}), q_\phi(\mathbf{u}\|\mathbf{z}, \mathbf{x}), q_\phi(\mathbf{v}\|\mathbf{z}, \mathbf{y}), p_{\mathsf{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency
2. Replace $I_{\mathsf{xy}\rightarrow}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$ with $I_{\mathsf{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$

Derivation

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$:

minimize	$I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$D(p_{\text{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v}), q_{\text{xy}\to}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})) = 0$
variables	$q_\phi(\mathbf{z}\|\mathbf{x}, \mathbf{y}), q_\phi(\mathbf{u}\|\mathbf{z}, \mathbf{x}), q_\phi(\mathbf{v}\|\mathbf{z}, \mathbf{y}), p_{\text{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency
2. Replace $I_{\text{xy}\to}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$ with $I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$

Derivation

- For each model $p_{\mathsf{model}} \in \{p_{\to\mathsf{xy}}, p_{\mathsf{x}\to\mathsf{y}}, p_{\mathsf{y}\to\mathsf{x}}\}$:

minimize	$I_{\mathsf{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$D(p_{\mathsf{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v}), q_{\mathsf{xy}\to}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})) = 0$
variables	$q_\phi(\mathbf{z}\|\mathbf{x}, \mathbf{y}), q_\phi(\mathbf{u}\|\mathbf{z}, \mathbf{x}), q_\phi(\mathbf{v}\|\mathbf{z}, \mathbf{y}), p_{\mathsf{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency
2. Replace $I_{\mathsf{xy}\to}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$ with $I_{\mathsf{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
3. Relax the equality constraint

Derivation

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$:

minimize	$I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
subject to	$D(p_{\text{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v}), q_{\text{xy}\to}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})) \le \epsilon$
variables	$q_\phi(\mathbf{z}\|\mathbf{x}, \mathbf{y}), q_\phi(\mathbf{u}\|\mathbf{z}, \mathbf{x}), q_\phi(\mathbf{v}\|\mathbf{z}, \mathbf{y}), p_{\text{model}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency
2. Replace $I_{\text{xy}\to}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$ with $I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$
3. Relax the equality constraint

Derivation

- For each model $p_{\text{model}} \in \{p_{\to \text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$:

minimize	$I_{\text{model}}(\mathbf{X},\mathbf{Y};\mathbf{Z})$
subject to	$D(p_{\text{model}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}), q_{\text{xy}\to}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})) \le \epsilon$
variables	$q_\phi(\mathbf{z}\vert\mathbf{x},\mathbf{y}), q_\phi(\mathbf{u}\vert\mathbf{z},\mathbf{x}), q_\phi(\mathbf{v}\vert\mathbf{z},\mathbf{y}), p_{\text{model}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})$

1. Replace $\mathbf{X}-\mathbf{Z}-\mathbf{Y}$ with the model consistency
2. Replace $I_{\text{xy}\to}(\mathbf{X},\mathbf{Y};\mathbf{Z})$ with $I_{\text{model}}(\mathbf{X},\mathbf{Y};\mathbf{Z})$
3. Relax the equality constraint
4. Convert to an unconstrained Lagrangian minimization

Derivation

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$:

minimize	$D(p_{\text{model}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v}), q_{\text{xy}\to}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})) + \lambda^{\text{CI}}_{\text{model}} I_{\text{model}}(\mathbf{X},\mathbf{Y};\mathbf{Z})$
subject to	
variables	$q_\phi(\mathbf{z}\|\mathbf{x},\mathbf{y}), q_\phi(\mathbf{u}\|\mathbf{z},\mathbf{x}), q_\phi(\mathbf{v}\|\mathbf{z},\mathbf{y}), p_{\text{model}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})$

1. Replace $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$ with the model consistency
2. Replace $I_{\text{xy}\to}(\mathbf{X},\mathbf{Y};\mathbf{Z})$ with $I_{\text{model}}(\mathbf{X},\mathbf{Y};\mathbf{Z})$
3. Relax the equality constraint
4. Convert to an unconstrained Lagrangian minimization

Training method

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$,

$$\boxed{\text{minimize} \quad D(p_{\text{model}}, q_{\text{xy}\to}) + \lambda_{\text{model}}^{\text{CI}} I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})}$$

- Distribution matching with CI regularization

Training method

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$,

$$\boxed{\text{minimize} \quad D(p_{\text{model}}, q_{\text{xy}\to}) + \lambda^{\text{CI}}_{\text{model}} I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})}$$

- Distribution matching with CI regularization
- Simultaneous training: minimize a weighted sum of the objectives

Training method

- For each model $p_{\text{model}} \in \{p_{\rightarrow\text{xy}}, p_{\text{x}\rightarrow\text{y}}, p_{\text{y}\rightarrow\text{x}}\}$,

$$\boxed{\text{minimize} \quad D(p_{\text{model}}, q_{\text{xy}\rightarrow}) + \lambda_{\text{model}}^{\text{CI}} I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})}$$

- Distribution matching with CI regularization
- Simultaneous training: minimize a weighted sum of the objectives
- Symmetric KL divergence $D_{\text{sym}}(p, q) \triangleq D_{\text{KL}}(p \parallel q) + D_{\text{KL}}(q \parallel p)$

Training method

- For each model $p_{\text{model}} \in \{p_{\to\text{xy}}, p_{\text{x}\to\text{y}}, p_{\text{y}\to\text{x}}\}$,

$$\boxed{\text{minimize} \quad D(p_{\text{model}}, q_{\text{xy}\to}) + \lambda_{\text{model}}^{\text{CI}} I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})}$$

- Distribution matching with CI regularization
- Simultaneous training: minimize a weighted sum of the objectives
- Symmetric KL divergence $D_{\text{sym}}(p, q) \triangleq D_{\text{KL}}(p \parallel q) + D_{\text{KL}}(q \parallel p)$
- Variational density-ratio estimation technique [Pu+17]: a variant of the discriminator trick of generative adversarial netwokrs (GANs)

Training method

- For each model $p_{\text{model}} \in \{p_{\rightarrow\text{xy}}, p_{\text{x}\rightarrow\text{y}}, p_{\text{y}\rightarrow\text{x}}\}$,

$$\boxed{\text{minimize} \quad D(p_{\text{model}}, q_{\text{xy}\rightarrow}) + \lambda_{\text{model}}^{\text{CI}} I_{\text{model}}(\mathbf{X}, \mathbf{Y}; \mathbf{Z})}$$

- Distribution matching with CI regularization
- Simultaneous training: minimize a weighted sum of the objectives
- Symmetric KL divergence $D_{\text{sym}}(p, q) \triangleq D_{\text{KL}}(p \parallel q) + D_{\text{KL}}(q \parallel p)$
- Variational density-ratio estimation technique [Pu+17]: a variant of the discriminator trick of generative adversarial netwokrs (GANs)
- In practice, $\lambda_{\text{model}}^{\text{CI}}$ can be chosen by trial-and-errors

Training with variational density-ratio estimation technique

- A variant of the discriminator trick of generative adversarial networks [Pu+17]

Training with variational density-ratio estimation technique

- A variant of the discriminator trick of generative adversarial networks [Pu+17]
- Suppose we wish to solve

$$\min_\theta D_{\mathsf{KL}}(p_\theta(s) \parallel q_\theta(s)) = \mathsf{E}_{p_\theta(s)}\left[\log \frac{p_\theta(s)}{q_\theta(s)}\right]$$

with no explicit density models on $p_\theta(s)$ and $q_\theta(s)$

Training with variational density-ratio estimation technique

- A variant of the discriminator trick of generative adversarial networks [Pu+17]
- Suppose we wish to solve

$$\min_{\theta} D_{\mathsf{KL}}(p_\theta(s) \parallel q_\theta(s)) = \mathsf{E}_{p_\theta(s)}\left[\log \frac{p_\theta(s)}{q_\theta(s)}\right]$$

  with no explicit density models on $p_\theta(s)$ and $q_\theta(s)$
- We can approximate the objective if we knew how to compute the ratio $p_\theta(s)/q_\theta(s)$

Training with variational density-ratio estimation technique

- A variant of the discriminator trick of generative adversarial networks [Pu+17]
- Suppose we wish to solve

$$\min_{\theta} D_{\mathsf{KL}}(p_\theta(s) \parallel q_\theta(s)) = \mathsf{E}_{p_\theta(s)}\left[\log \frac{p_\theta(s)}{q_\theta(s)}\right]$$

  with no explicit density models on $p_\theta(s)$ and $q_\theta(s)$
- We can approximate the objective if we knew how to compute the ratio $p_\theta(s)/q_\theta(s)$
- **Variational density-ratio estimation**: We learn $r_\psi(s) \approx p_\theta(s)/q_\theta(s)$ by solving

$$D_{\mathsf{JS}}(p_\theta(s), q_\theta(s)) = \max_{\psi} \mathsf{E}_{p_\theta(s)}[\log \sigma(\log r_\psi(s))] + \mathsf{E}_{q_\theta(s)}[\log \sigma(-\log r_\psi(s))]$$

Training with variational density-ratio estimation technique

- A variant of the discriminator trick of generative adversarial networks [Pu+17]
- Suppose we wish to solve

$$\min_\theta D_{\mathsf{KL}}(p_\theta(s) \,\|\, q_\theta(s)) = \mathsf{E}_{p_\theta(s)}\left[\log \frac{p_\theta(s)}{q_\theta(s)}\right]$$

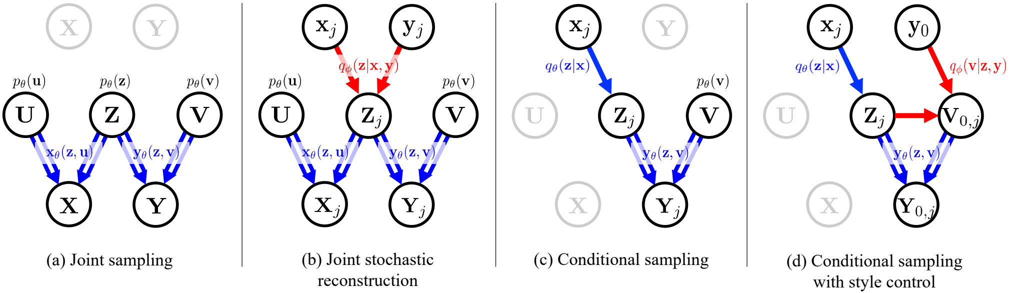
  with no explicit density models on $p_\theta(s)$ and $q_\theta(s)$
- We can approximate the objective if we knew how to compute the ratio $p_\theta(s)/q_\theta(s)$
- **Variational density-ratio estimation**: We learn $r_\psi(s) \approx p_\theta(s)/q_\theta(s)$ by solving

$$D_{\mathsf{JS}}(p_\theta(s), q_\theta(s)) = \max_\psi \mathsf{E}_{p_\theta(s)}[\log \sigma(\log r_\psi(s))] + \mathsf{E}_{q_\theta(s)}[\log \sigma(-\log r_\psi(s))]$$

- **Plug in and optimize**:

$$\min_\theta D_{\mathsf{KL}}(p_\theta(s) \,\|\, q_\theta(s)) \approx \min_\theta \mathsf{E}_{p_\theta(s)}[\log r_\psi(s)]$$

How to use the variational Wyner model



- Variational encoders are introduced for training, but can be also used in sampling
- Local variational encoders $q_\phi(\mathbf{u}|\mathbf{z},\mathbf{x}), q_\phi(\mathbf{v}|\mathbf{z},\mathbf{y})$ can be viewed as style extractors

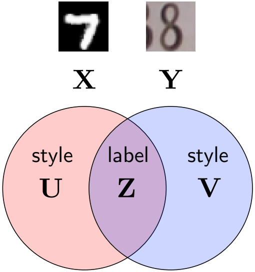
Experiment. MNIST–SVHN add-1 dataset

- $(\mathbf{X}, \mathbf{Y})$=(MNIST, SVHN) with label(SVHN)=label(MNIST)+1

Experiment. MNIST–SVHN add-1 dataset

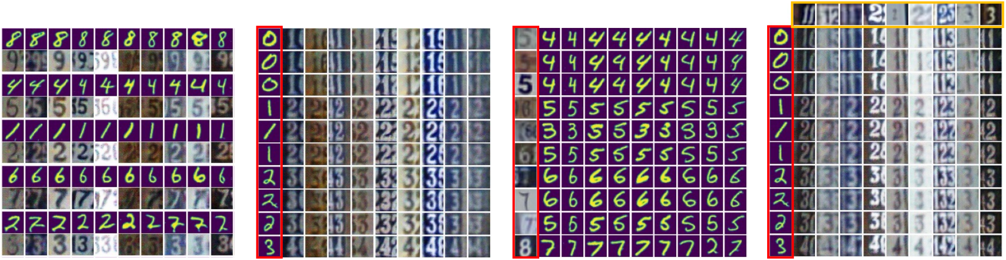
- $(\mathbf{X}, \mathbf{Y})$=(MNIST, SVHN) with label(SVHN)=label(MNIST)+1

- $\mathbf{Z}$=label, $(\mathbf{U}, \mathbf{V}) \approx$ (style of MNIST, style of SVHN)



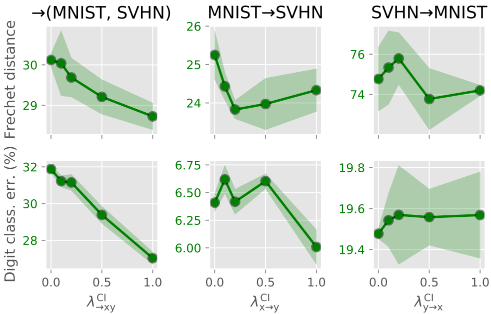
Experiment. MNIST–SVHN add-1 dataset

- Generated samples: same labels across the rows; same styles across the columns
- A red box highlights inputs; a yellow box highlight style references



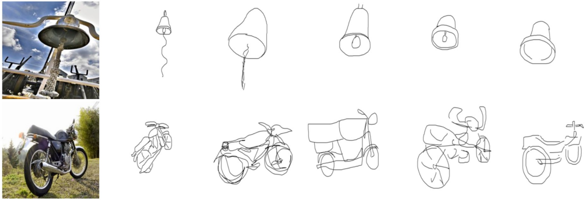
(a) →(MNIST,SVHN) (b) MNIST→SVHN (c) SVHN→MNIST (d) MNIST→SVHN with style transfer

- Numerical evaluation: $\lambda^{\mathrm{CI}}_{\mathrm{model}}$ vs. quality of generated samples



Experiment. Sketchy dataset [San+16]

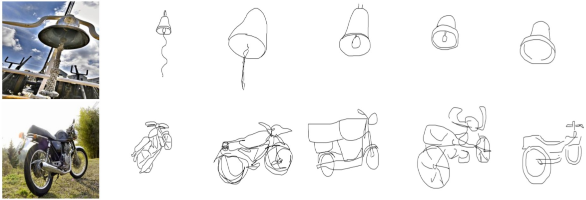
- $(\mathbf{X}, \mathbf{Y})$=(photo, human sketch)



- $\mathbf{Z} \approx$ image class, $(\mathbf{U}, \mathbf{V}) \approx$ (variation in photo, style of sketch)

Experiment. Sketchy dataset [San+16]

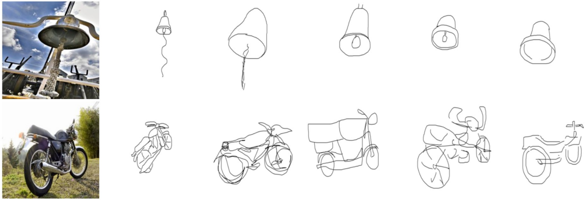
- $(\mathbf{X}, \mathbf{Y})$=(photo, human sketch)



- $\mathbf{Z} \approx$ image class, $(\mathbf{U}, \mathbf{V}) \approx$ (variation in photo, style of sketch)
- Cross-domain retrieval: given a sketch $(\mathbf{y})$, retrieve photos $(\mathbf{x})$

Experiment. Sketchy dataset [San+16]

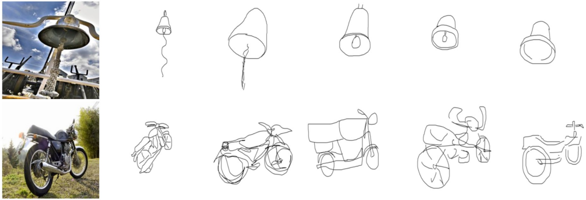
- $(\mathbf{X}, \mathbf{Y})$=(photo, human sketch)



- $\mathbf{Z} \approx$ image class, $(\mathbf{U}, \mathbf{V}) \approx$ (variation in photo, style of sketch)
- Cross-domain retrieval: given a sketch $(\mathbf{y})$, retrieve photos $(\mathbf{x})$
- Zero-shot: training set has no overlapping classes with test set

Experiment. Sketchy dataset [San+16]

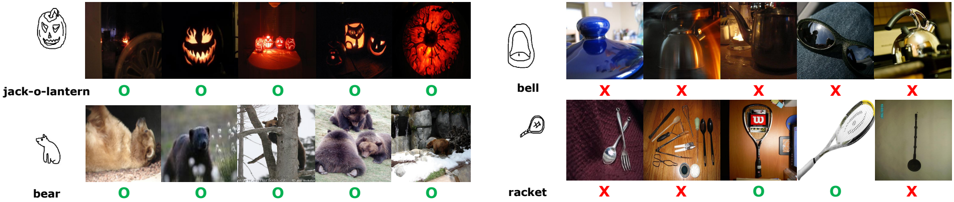
- $(\mathbf{X}, \mathbf{Y})$=(photo, human sketch)



- $\mathbf{Z} \approx$ image class, $(\mathbf{U}, \mathbf{V}) \approx$ (variation in photo, style of sketch)
- Cross-domain retrieval: given a sketch $(\mathbf{y})$, retrieve photos $(\mathbf{x})$
- Zero-shot: training set has no overlapping classes with test set
- Our method: retrieve via common representations

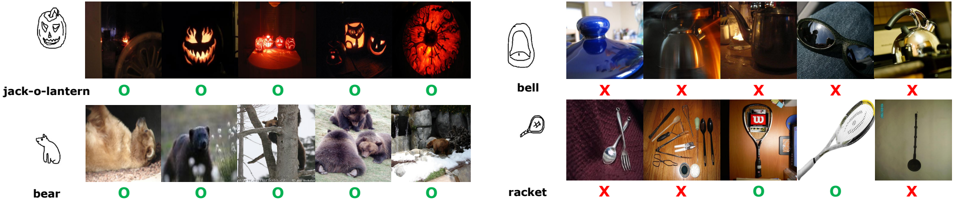
Experiment. Sketchy dataset [San+16]

- **Examples**: correct (left)/wrong (right) retrievals



Experiment. Sketchy dataset [San+16]

- **Examples**: correct (left)/wrong (right) retrievals



- Numerical evaluation: precision@K (P@K), mean average precision (mAP)

Models	P@100	mAP
LCALE [Lin+20]	0.583	0.476
IIAE [Hwa+20]	0.659	0.573
Variational Wyner	**0.703**	**0.629**

Experiment. CUB image-caption

- $(\mathbf{X}, \mathbf{Y})$=(bird images, captions)

the bird has a white body, black wings, and webbed orange feet

a blue bird with gray primaries and secondaries and white breast and throat

- Used ResNet-101 features for images

Experiment. CUB image-caption

→(image, caption)

this small bird is black white white with a small bill bill and black feet	this bird is grey with grey and a black beak , pointy short pointy beak .	this is a black and white black bird and a short black beak .	this bird has a black and and white and white feathers and

this white bird is mostly white white with a long bill , and black feet	this bird is grey with grey and has long long, pointy short pointy beak	this is a black and white black bird and a long long yellow . .	this bird has a white and and white and white with and feet .

image→caption

input image from test set	generated captions		
	this bird has a black crown and breast , with a crown , and and and black red its . .	this is a very , and white and and color with with a , and and a long blue patches . .	this bird has a very , thin beak with a breast and a brown beak , the body rimmed body .
	this bird has a black crown and breast , with yellow breast and and and its of its feathers .	this bird a small , and yellow black color with with crown black black and black of its crown .	this bird has yellow small , black beak and a breast and a black feathers . the bird it's the body . .
	this bird has a red crown and breast , with red red red and red and on on red its . .	this bird is a red red , red red color with with crown red and and and black red red .	the bird has red red red red , and red red and a red beak . the red 's feathers .

caption→image

input text from test set	ground truth / retrievals from generated features
This bird has yellow topped black and white striped wings and some red markings on its belly.	
This bird has wings that are gray and has a white belly.	

Experiment. CUB image-caption

- Numerical evaluation: correlation of generated samples

Model	joint	image→caption	caption→image
Test set		0.273	
MMVAE [Shi+19]	0.263	0.104	0.135
Variational Wyner	**0.303**	**0.327**	**0.318**

Concluding remarks

- Learning distributions with Wyner's common information
  $\rightarrow$ disentangled representations + better performance in downstream tasks!

Concluding remarks

- Learning distributions with Wyner's common information
  $\rightarrow$ disentangled representations + better performance in downstream tasks!
- Wyner's common information J: $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$

$$\min_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} I(\mathbf{Z};\mathbf{X},\mathbf{Y})$$

Concluding remarks

- Learning distributions with Wyner's common information
  $\to$ disentangled representations + better performance in downstream tasks!
- Wyner's common information J: $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$

$$\min_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} I(\mathbf{Z};\mathbf{X},\mathbf{Y})$$

Concluding remarks

- Learning distributions with Wyner's common information
  $\rightarrow$ disentangled representations + better performance in downstream tasks!
- Wyner's common information J: $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$

$$\min_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} I(\mathbf{Z};\mathbf{X},\mathbf{Y})$$

- Is this the end of the story?

Concluding remarks

- Learning distributions with Wyner's common information
  $\rightarrow$ disentangled representations + better performance in downstream tasks!
- Wyner's common information J: $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$

$$\min_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} I(\mathbf{Z};\mathbf{X},\mathbf{Y})$$

- Is this the end of the story?

Q1 What is the operational meaning of learning representation based on Wyner's optimization problem?

Concluding remarks

- Learning distributions with Wyner's common information
  $\rightarrow$ disentangled representations + better performance in downstream tasks!
- Wyner's common information J: $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$

$$\min_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} I(\mathbf{Z};\mathbf{X},\mathbf{Y})$$

- Is this the end of the story?

Q1 What is the operational meaning of learning representation based on Wyner's optimization problem?

Q2 More than two variables?

Concluding remarks

- Learning distributions with Wyner's common information
  $\to$ disentangled representations + better performance in downstream tasks!
- Wyner's common information J: $\mathbf{X} - \mathbf{Z} - \mathbf{Y}$

$$\min_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} I(\mathbf{Z};\mathbf{X},\mathbf{Y})$$

- Is this the end of the story?

Q1 What is the operational meaning of learning representation based on Wyner's optimization problem?

Q2 More than two variables?

"Learning with Succinct Common Representation with Wyner's Common Information", **J. Jon Ryu**, Yoojin Choi, Young-Han Kim, Mostafa El-Khamy, and Jungwon Lee, arXiv:1905.10945, Under Review

References I

[Hwa+20] H. Hwang, G.-H. Kim, S. Hong, and K.-E. Kim. "Variational Interaction Information Maximization for Cross-domain Disentanglement". In: *Adv. Neural Info. Proc. Syst.* Vol. 33. 2020.

[Lin+20] K. Lin, X. Xu, L. Gao, Z. Wang, and H. T. Shen. "Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval". In: *Proc. AAAI Conf. Artif. Int.* Vol. 34. 2020, pp. 11515–11522.

[Pu+17] Y. Pu, W. Wang, R. Henao, L. Chen, Z. Gan, C. Li, and L. Carin. "Adversarial symmetric variational autoencoder". In: *Adv. Neural Info. Proc. Syst.* 2017, pp. 4330–4339.

[San+16] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. "The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies". In: *ACM Trans. Graph. (Proc. SIGGRAPH)* (2016).

[Shi+19] Y. Shi, N. Siddharth, B. Paige, and P. H. Torr. "Variational mixture-of-experts autoencoders for multi-modal deep generative models". In: *Adv. Neural Info. Proc. Syst.* Vol. 32. 2019.