The Variational Fair Autoencoder

Christos Louizos, Kevin Swersky, Yujia Li
Max Welling, Richard Zemel

CIFAR
CANADIAN INSTITUTE
for ADVANCED RESEARCH

Motivation

- Consider the task of identifying a person in the following images:

- Can be hard since a lot of “noise” obfuscates the predictive information

Motivation (2)

- Determine possible suspects from photos
  - Sensitive information (e.g. race and gender) of the individual should not affect decisions
- Detect Alzheimer on MRI images
  - MRI images from machine 1 and 2
  - Avoid machine related variations for better generalization

Tackling such problems

- Simply excluding these particular bits from the input is not going to work
  - Other dimensions still contain information about these bits
- Transform the data to a new *representation*
  - Explicitly encode its properties
    - Enforce invariance w.r.t. a-priori known information

Related work

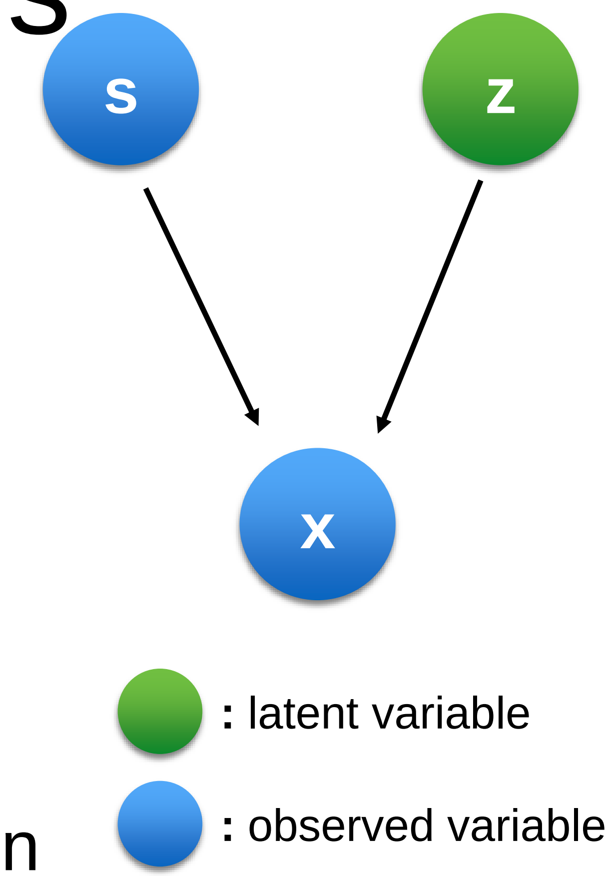
- "Learning Fair Representations"[1] (LFR)
  - Simple discriminative clustering approach
- Neural networks with a Maximum Mean Discrepancy[5] penalty[2, 7, 8]
- "Domain Adversarial Neural Networks"[3] (DANN)
  - A minimax problem

[1]"Learning Fair Representations",Zemel et al., 2013
[5]"A Kernel Two-Sample Test", Gretton et al, 2012
[2]"Learning unbiased features", Li et al., 2014
[7]"Deep Domain Confusion: Maximizing for Domain Invariance", Tzeng et al., 2014
[8]"Learning Transferable Features with Deep Adaptation Networks", Long et al., 2015
[3]Domain Adversarial Training of Neural Networks", Ganin et al., 2015

Contribution

- Variational Fair Autoencoder (VFAE)
  - A generative model where known/target factors of variation are explicitly removed
    - New representation is *invariant* w.r.t. this information
  - Better performance on fair classification, domain adaptation and general feature learning tasks

Unsupervised Variational Autoencoder[4] for invariant representations



- Two independent factors of variation
  - **s** : observed (discrete) "sensitive"/"nuisance" factors of variation
  - **z** : continuous latent variable for the remaining information

$p_\theta(\mathbf{x},\mathbf{z}|\mathbf{s}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z},\mathbf{s})$ as a neural network generative model (decoder)

$q_\varphi(\mathbf{z}|\mathbf{x},\mathbf{s})$ as a neural network variational posterior (encoder) since exact inference is intractable

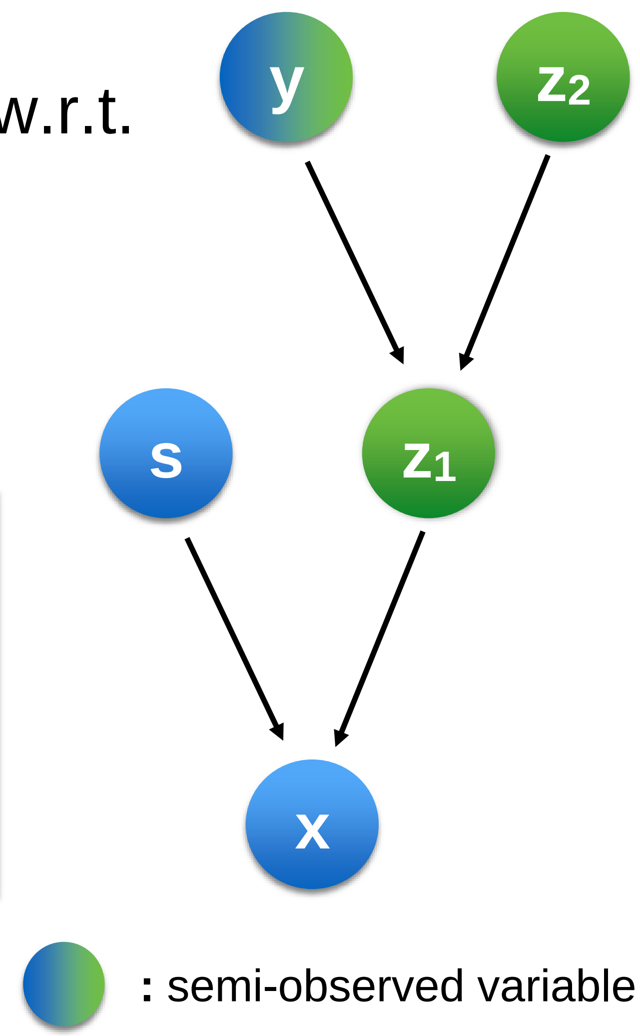
Objective Function

$$\sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{s}_n) \geq \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_n|\mathbf{x}_n,\mathbf{s}_n)}[\log p_\theta(\mathbf{x}_n|\mathbf{z}_n,\mathbf{s}_n)] - KL(q_\phi(\mathbf{z}_n|\mathbf{x}_n,\mathbf{s}_n)||p(\mathbf{z}))$$

[4]"Semi-Supervised Learning with Deep Generative Models", Kingma et al., 2014

Semi-Supervised VAE[4] for invariant representations

- Unsupervised model may create degenerate representations w.r.t. the prediction task (**y**)
- Enrich generative model so as to correlate **z** with **y**



$p_\theta(\mathbf{z_1},\mathbf{z_2},\mathbf{x},\mathbf{y}|\mathbf{s}) = p(\mathbf{y})p(\mathbf{z_2})p_\theta(\mathbf{z_1}|\mathbf{z_2},\mathbf{y})p_\theta(\mathbf{x}|\mathbf{z_1},\mathbf{s})$, as a neural network generative model (decoder)

$q_\varphi(\mathbf{z_1},\mathbf{z_2},\mathbf{y}|\mathbf{x},\mathbf{s}) = q_\varphi(\mathbf{z_1}|\mathbf{x},\mathbf{s})q_\varphi(\mathbf{y}|\mathbf{z_1})q_\varphi(\mathbf{z_2}|\mathbf{z_1},\mathbf{y})$ as a neural network variational posterior (encoder)

VAE Objective Function

$$\mathcal{F}_{\text{VAE}}(\phi,\theta;\mathbf{x}_n,\mathbf{x}_m,\mathbf{s}_n,\mathbf{s}_m,\mathbf{y}_n) = \sum_{n=1}^{N}\mathcal{L}_s(\phi,\theta;\mathbf{x}_n,\mathbf{s}_n,\mathbf{y}_n) + \sum_{m=1}^{M}\mathcal{L}_u(\phi,\theta;\mathbf{x}_m,\mathbf{s}_m) + \alpha\sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{z}_{1n}|\mathbf{x}_n,\mathbf{s}_n)}[-\log q_\phi(\mathbf{y}_n|\mathbf{z}_{1n})]$$

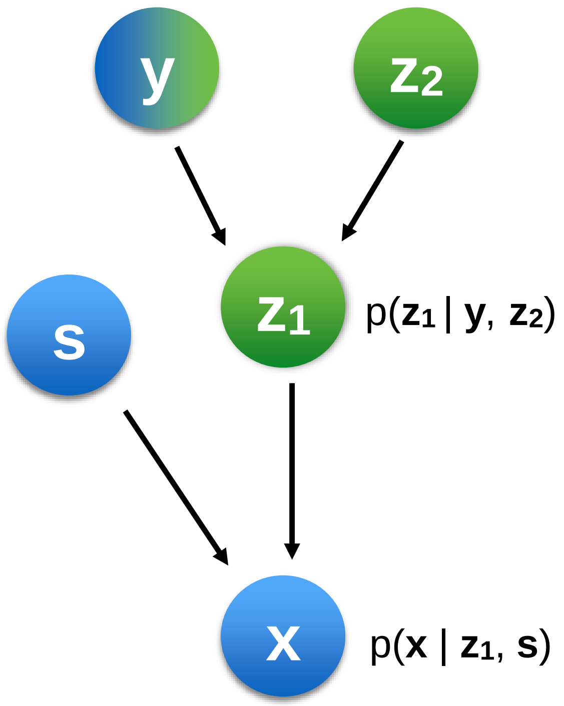
Further invariance via posterior regularization

- Model encourages independence between $\mathbf{z}_1$ and $\mathbf{s}$ a-priori

- Some dependencies might still remain in the (approximate) posterior $q(\mathbf{z}_1|\mathbf{s})$

  - e.g. if **s** and **y** are correlated then $q(\mathbf{y}|\mathbf{z}_1)$ can "leak" information about **s**

- Introduce an extra penalty term to avoid information about **s** as much as possible



Independence between $\mathbf{z}_1$ and $\mathbf{s}$ a-priori

Maximum Mean Discrepancy[5] (MMD)

- MMD measures the “distance” between two sets of samples

$$\ell_{\text{MMD}}(\mathbf{X}, \mathbf{X}') = \left\| \frac{1}{N_0} \sum_{i=1}^{N_0} \psi(\mathbf{x}_i) - \frac{1}{N_1} \sum_{j=1}^{N_1} \psi(\mathbf{x}'_j) \right\|^2$$

$$= \frac{1}{N_0^2} \sum_{i=1}^{N_0} \sum_{i'=1}^{N_0} k(\mathbf{x}_i, \mathbf{x}_{i'}) + \frac{1}{N_1^2} \sum_{j=1}^{N_1} \sum_{j'=1}^{N_1} k(\mathbf{x}'_j, \mathbf{x}'_{j'}) - \frac{2}{N_0 N_1} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} k(\mathbf{x}_i, \mathbf{x}'_j)$$

- For universal kernels (e.g. rbf) it is asymptotically 0 if both sample sets are “drawn” from the same distribution

[5]“A Kernel Two-Sample Test”, Gretton et al, 2012

Fast MMD via Random Fourier Features

- Computing MMD is expensive
  - Scales quadratically with the mini-batch size due to the Gram matrix
- Random Kitchen Sinks to approximate the rbf MMD[6]
  - Work with primal space: $\left\| \frac{1}{N_0} \sum_{i=1}^{N_0} \psi(\mathbf{x}_i) - \frac{1}{N_1} \sum_{i=1}^{N_1} \psi(\mathbf{x}_i') \right\|^2$
  - Scales linearly with the mini-batch size
- Feature expansion is given by:

$$\psi(\mathbf{x}) = \sqrt{\frac{2}{D}} \cos\left(\sqrt{\frac{2}{\gamma}} \mathbf{x}\mathbf{W} + \mathbf{b} \right)$$

$$\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}); \qquad \mathbf{b} \sim \mathcal{U}[0, 2\pi]$$

[6]"FastMMD: Ensemble of Circular Discrepancy for Efficient Two-Sample Test", Zhao et al., 2014

Variational Fair Autoencoder (VFAE)

- We incorporate MMD in the lower bound of our VAE
  - We split the samples from $q(\mathbf{z_1}|\mathbf{x},\mathbf{s})$ according to the state of $\mathbf{s}$
  - We treat those as samples from the marginal posteriors $q(\mathbf{z_1}|\mathbf{s})$

VFAE Objective Function

$$\mathcal{F}_{\text{VFAE}}(\phi,\theta;\mathbf{x}_n,\mathbf{x}_m,\mathbf{s}_n,\mathbf{s}_m,\mathbf{y}_n) = \mathcal{F}_{\text{VAE}}(\phi,\theta;\mathbf{x}_n,\mathbf{x}_m,\mathbf{s}_n,\mathbf{s}_m,\mathbf{y}_n) - \beta\ell_{\text{MMD}}(\mathbf{Z}_{1_{\mathbf{s}=0}},\mathbf{Z}_{1_{\mathbf{s}=1}})$$

$$\ell_{\text{MMD}}(\mathbf{Z}_{1_{\mathbf{s}=0}},\mathbf{Z}_{1_{\mathbf{s}=1}}) = \|\,\mathbb{E}_{\tilde{p}(\mathbf{x}|\mathbf{s}=0)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=0)}[\psi(\mathbf{z}_1)]] - E_{\tilde{p}(\mathbf{x}|\mathbf{s}=1)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=1)}[\psi(\mathbf{z}_1)]]\|^2$$

Experiments

1. Fair classification
2. Domain Adaptation
3. General feature learning

Evaluation criteria

- $\mathbf{z_1}$ should provide low (random chance) accuracy on **s** and high accuracy on **y**
  - Measured linearly (Logistic Regression) and non-linearly (Random Forest)
- $\mathbf{z_1}$ should also not "discriminate" for fair classification[1]
  - Ensure unbiased decisions from the classifier

$$\text{Discrimination}(\mathbf{y}_{s=0}, \mathbf{y}_{s=1}) = \left| \frac{\sum_{n=1}^{N} \mathbb{I}[y_n^{s=0}]}{N_{s=0}} - \frac{\sum_{n=1}^{N} \mathbb{I}[y_n^{s=1}]}{N_{s=1}} \right|$$

[1]"Learning Fair Representations",Zemel et al., 2013

Experiments

1. Fair classification
2. Domain Adaptation
3. General feature learning

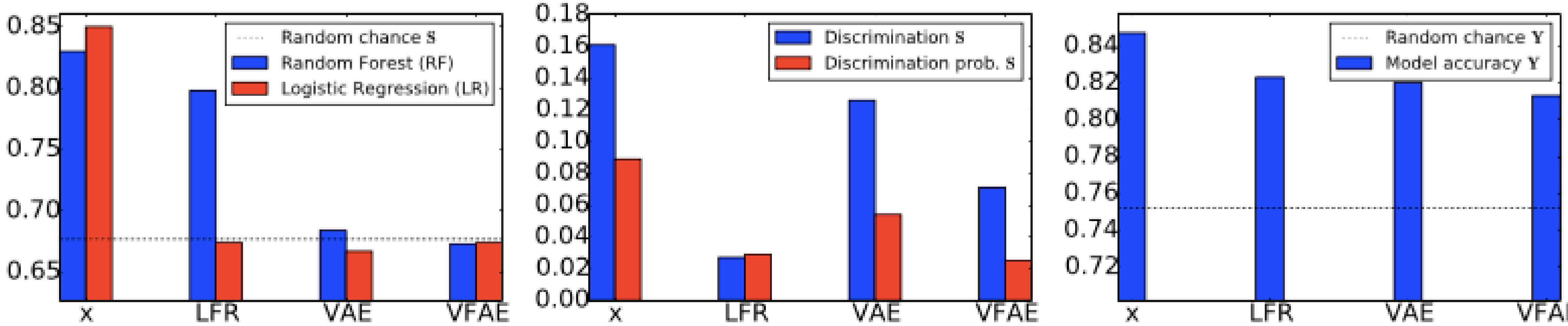
Fair Classification

- Adult dataset
  - **y**: account > 50.000$, **s**: gender
- Health dataset
  - **y**: whether admitted to hospital, **s**: age
- Learning Fair Representations[1] (LFR) as baseline

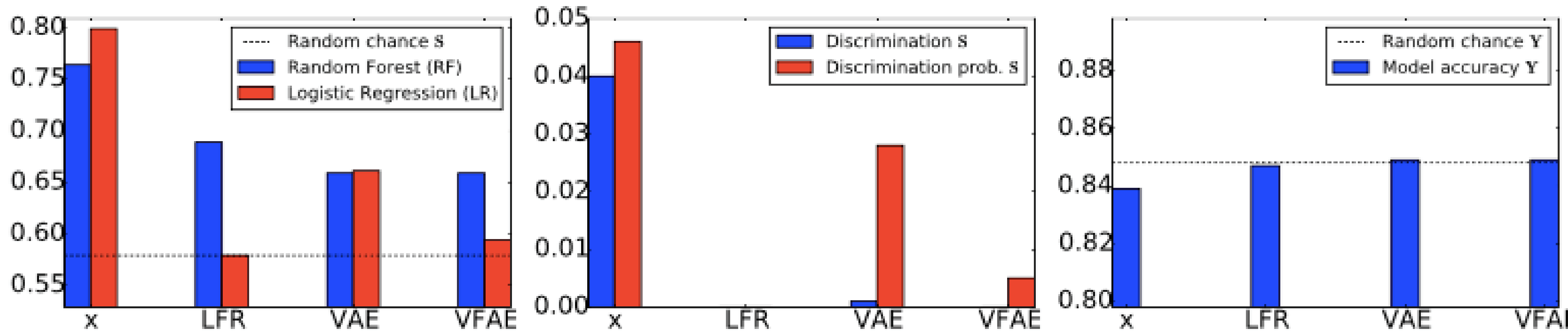
[1]“Learning Fair Representations”,Zemel et al., 2013

Fair classification results

Accuracy **S** | Discrimination **S** | Accuracy **y**



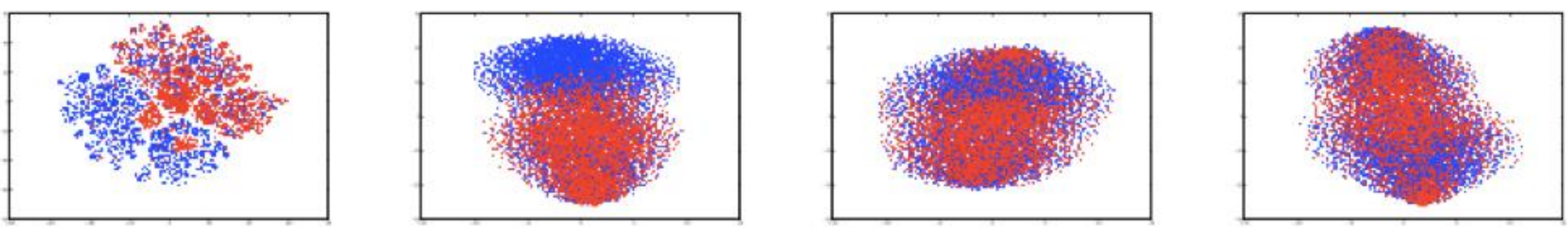
(a) Adult dataset



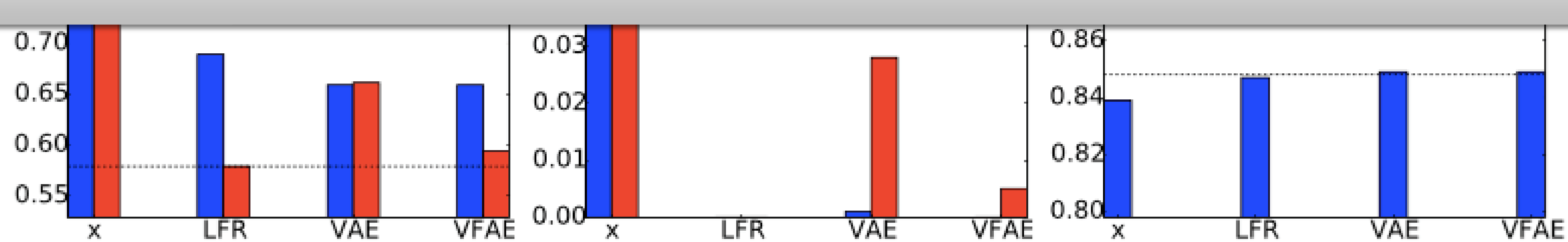
(c) Health dataset

Fair classification results

Accuracy **S** | Discrimination **S** | Accuracy **y**



(**A**): original **x**, (**B**): latent $\mathbf{z_1}$ without **s** and MMD, (**C**): latent $\mathbf{z_1}$ with **s** and without MMD, (**D**): latent $\mathbf{z_1}$ with **s** and MMD.



(c) Health dataset

Experiments

1. Fair classification
2. Domain Adaptation
3. General feature learning

Domain Adaptation

- Amazon reviews dataset
  - **y**: positive/negative review
  - **s**: domain (books, dvd, electronics, kitchen)
- Domain Adversarial Neural Networks[3] (DANN) as baseline

[3]Domain Adversarial Training of Neural Networks", Ganin et al., 2015

Domain adaptation results

Source - Target	S		Y	
	RF	LR	VFAE	DANN
books - dvd	0.535	0.564	**0.799**	0.784
books - electronics	0.541	0.562	**0.792**	0.733
books - kitchen	0.537	0.583	**0.816**	0.779
dvd - books	0.537	0.563	**0.755**	0.723
dvd - electronics	0.538	0.566	**0.786**	0.754
dvd - kitchen	0.543	0.589	**0.822**	0.783
electronics - books	0.562	0.590	**0.727**	0.713
electronics - dvd	0.556	0.586	**0.765**	0.738
electronics - kitchen	0.536	0.570	0.850	**0.854**
kitchen - books	0.560	0.593	**0.720**	0.709
kitchen - dvd	0.561	0.599	0.733	**0.740**
kitchen - electronics	0.533	0.565	0.838	**0.843**

Experiments

1. Fair classification

2. Domain Adaptation

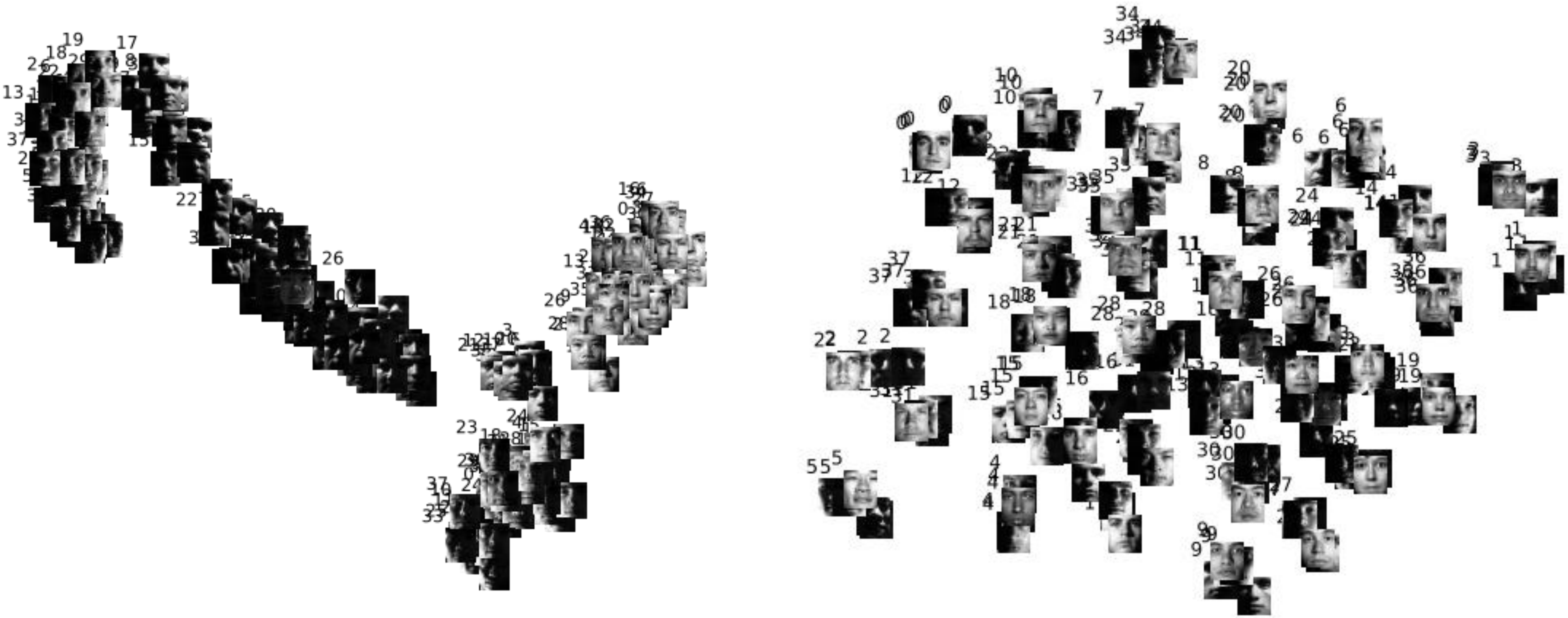
3. General feature learning

Invariant feature learning

- Extended Yale B dataset
  - Face images of 38 people under different lightning conditions
  - **y**: person ID
  - **s**: lightning condition of the photo
- A two hidden layer neural network with MMD[2] as the baseline

[2]"Learning unbiased features", Li et al., 2014

Invariant feature learning results



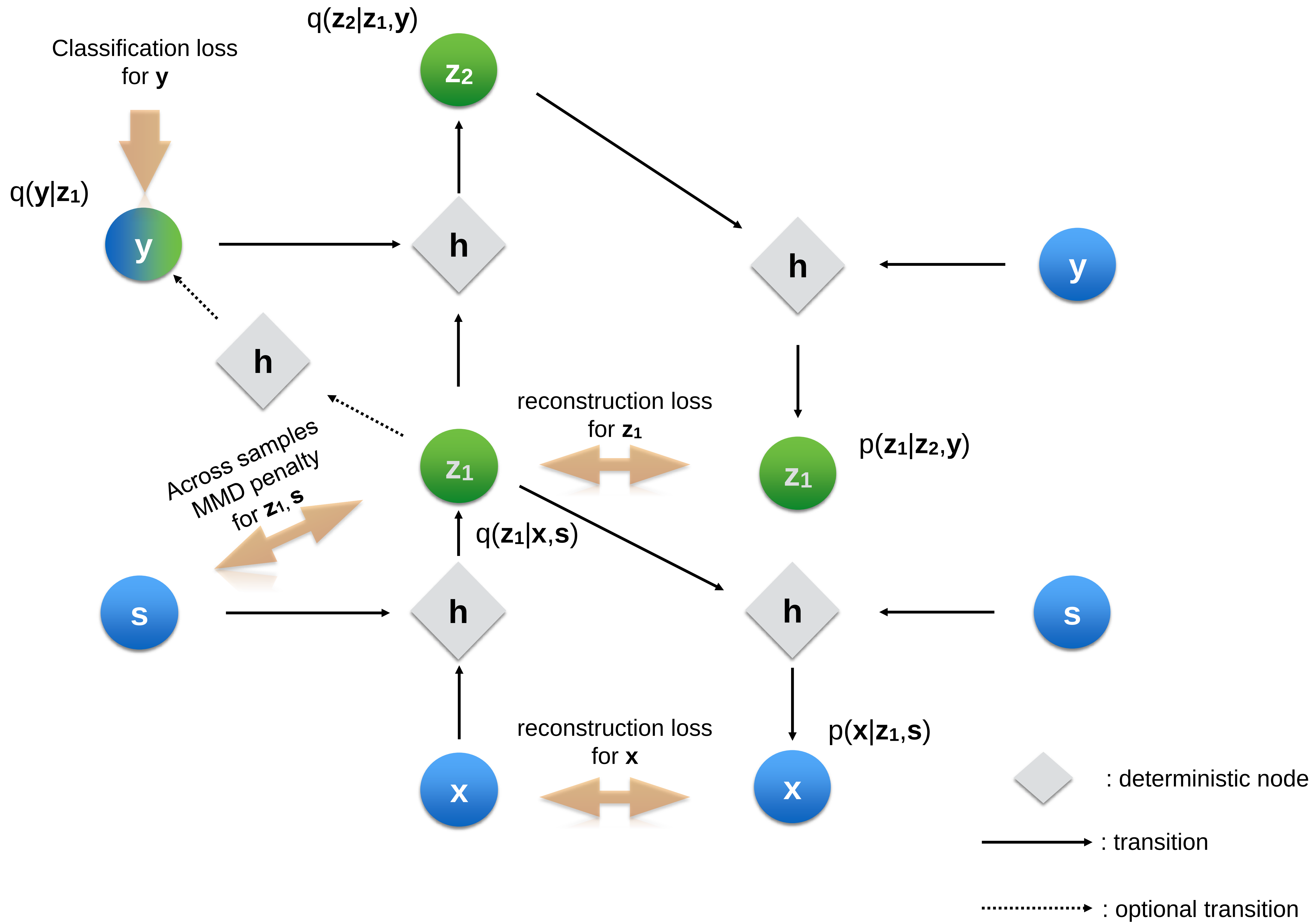
Method	S		Y
	RF	LR	
Original x	0.952	0.961	0.78
NN + MMD	-	-	0.82
VFAE	0.435	0.565	**0.846**

Conclusion & future work

- VFAE provides the better tradeoff in predicting **y** while obfuscating **s**
  - Incorporating MMD in VFAE helps
  - Effective in fair classification, domain adaptation and invariant feature learning
- Alternative posterior regularization techniques
  - Mutual information among the **s** and **z** distributions
- Extend to recommender systems
  - Recommendations that do not depend to sensitive demographic information

Thank you!

Questions?


q(z2|z1,y)
Classification loss
for y
z2
q(y|z1)
y
h
h
y
h
reconstruction loss
for z1
Across samples
MMD penalty
for z1, s
z1
z1
p(z1|z2,y)
q(z1|x,s)
s
h
h
s
reconstruction loss
for x
p(x|z1,s)
x
x
: deterministic node
: transition
: optional transition