Weekly Study Report

Wang Ma

2024-09-13

Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute

Outline

1. Paper Reading: Uncertainty in Causal Graphs

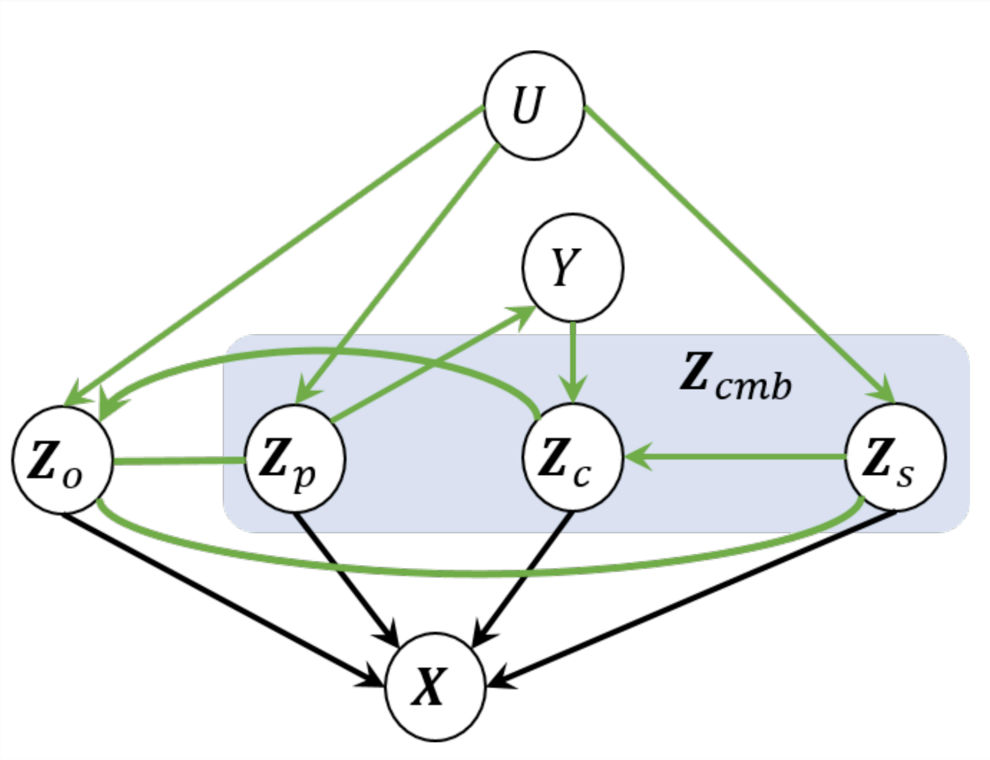
1.1 Causal Relationship Vs. Correlation

In real-world tasks, such as prediction, classification or decision-making, data are not only correlated with the target, the relationships are determined by latent causal relations.

- Traditional ML: focus on the **correlation** among variables
- Causal Inference: try to capture the **causal mechanisms** among variables, say, how does a variable influences our targets variable

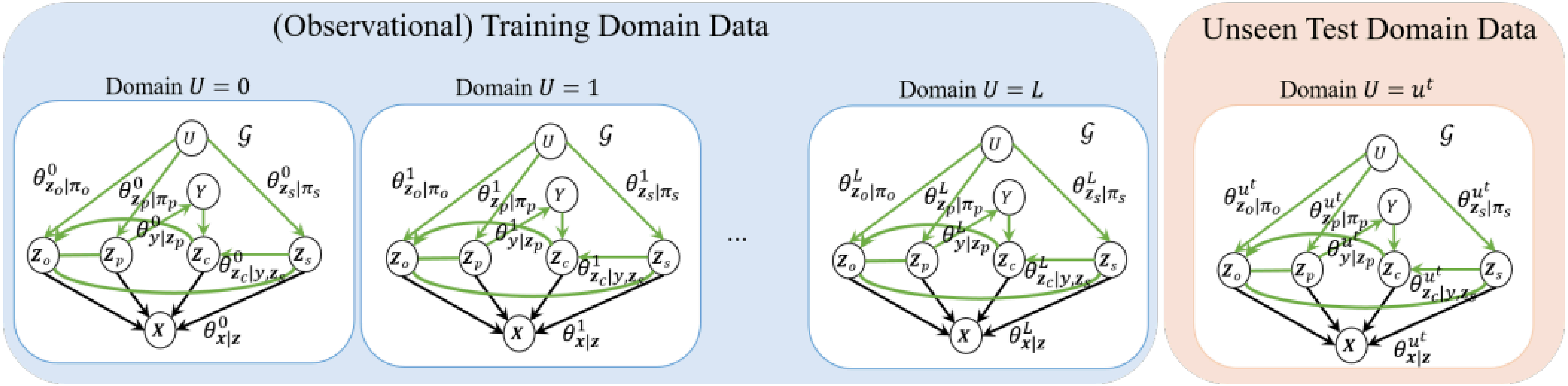
This helps us better **understand the data generation mechanism** and **make robust predictions under different conditions**, such as varying domains.

1.2.1 Data Generation Process



- X: high-dimensional data
- Y: target variable for prediction
- U: domain-specific information
- Z: latent, high-level variables for generation X
  - Z_p: **parent variables** which directly influence Y
  - Z_c: **child variables** directly affected by Y
  - Z_s: **spouse variables** related to Y through other connections
  - Z_o: **spurious variables** correlated with Y but not causally linked

1.2.2 Prediction Tasks uncer Domain Generalizations Settings



- **Prediction goal**: $p(y|x^t, D)$
- **MLE approach**: $G^* = \max p(D|G)$. Challenging, requiring a sufficient number of data, worse than SOTA domain generalization approaches.

1.2.3 Bayesian Inference: Sampling G from constructed posterior $p(G|D)$

The Bayesian causal discovery is usually employed when:

- the data is limited
- point-estimation causal discovery methods lead to poorly calibrated predictions.

More importantly, BI renders the ability to **quantify uncertainty**.

$$p(y|\boldsymbol{x}^t, \mathcal{D}) = \int_{\mathcal{G}} p(y|\boldsymbol{x}^t, \mathcal{G}, \mathcal{D}) p(\mathcal{G}|\boldsymbol{x}^t, \mathcal{D})\, \mathrm{d}\mathcal{G} \propto \mathbb{E}_{\mathcal{G}\sim p(\mathcal{G}|\mathcal{D})} \left[p(y|\boldsymbol{x}^t, \mathcal{G}) p(\boldsymbol{x}^t|\mathcal{G}) \right] \tag{1}$$

1.2.3.1 The Invariant Prediction Mechanism

- Z_{cmb}: the Causal Markov Blanket (CMB) variables, containing Z_p, Z_c and Z_s.

$$p(y|\boldsymbol{x}^t, \mathcal{G}) = \int_{\boldsymbol{z}} \sum_{u} p(y|\boldsymbol{x}^t, \boldsymbol{z}, u, \mathcal{G}) p(\boldsymbol{z}, u|\boldsymbol{x}^t, \mathcal{G}) \,\mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{z}^{\mathcal{G}}_{cmb}} p(y|\boldsymbol{z}^{\mathcal{G}}_{cmb}) p(\boldsymbol{z}^{\mathcal{G}}_{cmb}|\boldsymbol{x}^t, \mathcal{G}) \,\mathrm{d}\boldsymbol{z}^{\mathcal{G}}_{cmb}$$

1.2.3.2 Sample Density Estimation in Graphs

Recall Eq.(1):

$$p(y|x^t, D) \propto \mathbb{E}_{G \sim p(G|D)}[p(y|x^t, D) p(x^t|G)]$$

Directly get $p(x|G)$ is challenging due to the unavailability of U for x^t in the target domain; the causal mechanisms in the target domain are also unknown.

$$p(x^t \mid G) \propto e^{-\alpha U_e(x|G)}$$

1.2.4 Uncertainty Quantification in 3 Levels

1. **Causal Graph Uncertainty** $U(G)$: Quantifies the uncertainty in the causal graph's posterior distribution, indicating confidence in the learned graph. It can be calculated from $p(G|D)$.
2. **Single-Graph Prediction Uncertainty** $U_e(x|G)$: Measures uncertainty in predictions for a given graph G , which is critical for OOD predictions. It can be calculated from $p(y|x^t, G)$ (epistemic uncertainty).
3. **Bayesian Inference Uncertainty** $U(x|D)$: Quantifies the uncertainty in the final predictions by incorporating all possible graphs. It can be calculated from $p(y|x^t, D)$.

1.3.1 The Training Procedure

1. Learning Latent Variables via iVAE

$$\mathcal{L}_{\text{iVAE}} = \underbrace{-\mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) + \log p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|y,u) - \log q_\psi(\boldsymbol{z}|\boldsymbol{x})\right]}_{\mathcal{L}_{\text{ELBO}}} + \underbrace{\mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\nabla_{\boldsymbol{z}} q_\psi(\boldsymbol{z}|\boldsymbol{x}) - \nabla_{\boldsymbol{z}} p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|y,u)\right\|^2\right]}_{\mathcal{L}_{\text{SM}}}$$

2. Bayesian Causal Discovery via DAG-GFlowNet
   - DAG-GFLowNet: **estimates the posterior distribution** over causal graphs $p(G|D)$
   - Goal: The goal of this step is **to sample a diverse set of causal graphs** from the posterior distribution $p(G|D)$, capturing uncertainty about the true causal structure. These graphs are crucial for the Bayesian inference procedure.

3. Invariant Prediction Mechanism Learning
   - A prediction model $p_\varphi\left(Y|Z_{cmb}^{G_l}\right)$ is trained for each sampled causal graph G_l using the identified CMB variables.

1.3.2 The Inference Procedure

Obtained: a causal graph set $G = \{G^l\}_{l=1}^{L}$ and predictors $\left\{p_{\varphi^l}\left(y|Z_{cmb}^{G_l}\right)\right\}$. Given an input x^t from test domain, we

1. **Compute Single-Graph Prediction Uncertainty** $\{U_e(x^t|G^l)\}_{l=1}^{L}$

   To identify which causal graphs are more suitable for predicting a given test sample. This allows the model to weigh predictions from different graphs based on their fit to the new data.

2. **Estimate Data Density** $\{p(x^t|G^l)\}_{l=1}^{L}$

   The goal is to estimate the likelihood of the test sample under each causal graph to prioritize predictions from graphs that are more consistent with the test data.

3. **Bayesian Model Averaging for Final Prediction**

$$p(y|x^t, D) \propto \mathbb{E}_{G^l \sim p(G|D)}\left[p\left(y|z_{cmb}^{G_l}\right)p(x^t|G^l)\right]$$

2. Paper Reading: Diversity-enhanced Probabilistic Ensemble

Laplacian Approximation: to construct the posterior distribution $p(\theta \mid D, \beta)$ around a θ_{map}, where

$$\theta_{map} = \arg\max_{\theta} \log p(\theta \mid D, \beta).$$

And we have

$$p(\theta|D, \beta) \approx N\big(\theta_{map}, \Sigma\big),$$

where $\Sigma = -(H)^{-1}$ and $H = \nabla_\theta^2 \log p(\theta|D, \beta)|\theta = \theta_{map}$.

A mixture of Gaussian in constructed to better approximate the posterior distribution:

$$p(\theta|D,\beta) \approx \sum_{i=1}^{N} \lambda_i N(\theta;\theta_i,\Sigma_i).$$

The Bayesian predictive function:

$$p(y|x,D) \approx \int p(y|x,\theta) \sum_{i=1}^{N} \lambda_i N(\theta;\theta_i,\Sigma_i) d\theta$$

$$\approx \frac{1}{S} \sum_{i=1}^{S} p(y|x,\theta^s)$$

Proposition 3.1: Convergence of PE

$$\sup_{\theta} |p(\theta|D,\beta) - \sum_{i=1}^{N} \lambda_i N(\theta;\theta_i,\Sigma_i)| \to 0$$

Proposition 3.2: Better posterior approximation

$$KL(p(\theta \mid D, \beta) \| p_{PE}(\theta)) \leq \sum_{i=1}^{N} \lambda_i KL\big(p(\theta; D, \beta) \parallel p^i_{LA}(\theta)\big)$$

Proposition 3.3: Error Reduction and Diversity Measurement

$$-\log \mathbb{E}_\theta[p(y^*|x,\theta)] \leq \mathbb{E}_\theta[-\log p(y^*|x,\theta)] - \inf_\theta \frac{1}{2p(y^*|x,\theta)^2} \mathbb{V}_\theta[p(y^*|x,\theta)] \quad (7)$$

where $\inf_\theta \frac{1}{p(y^*|x,\theta)^2}$ *is bounded given* $p(y^*|x,\theta) \in [0,1]$ *and* $\mathbb{V}_\theta[p(y^*|x,\theta)]$ *is the variance of probabilistic ensemble model prediction.*

$$\mathbb{V}_\theta[p(y^*|x,\theta)] = \mathbb{E}_\theta[(p(y^*|x,\theta) - \mathbb{E}_\theta[p(y^*|x,\theta)])^2] \quad (8)$$

Proposition 3.4: Enhanced Diversity of PE

$$p_{DE} = \sum \lambda_i \delta(\theta, \theta_i) \sim (\mu_D, \Sigma_D)$$

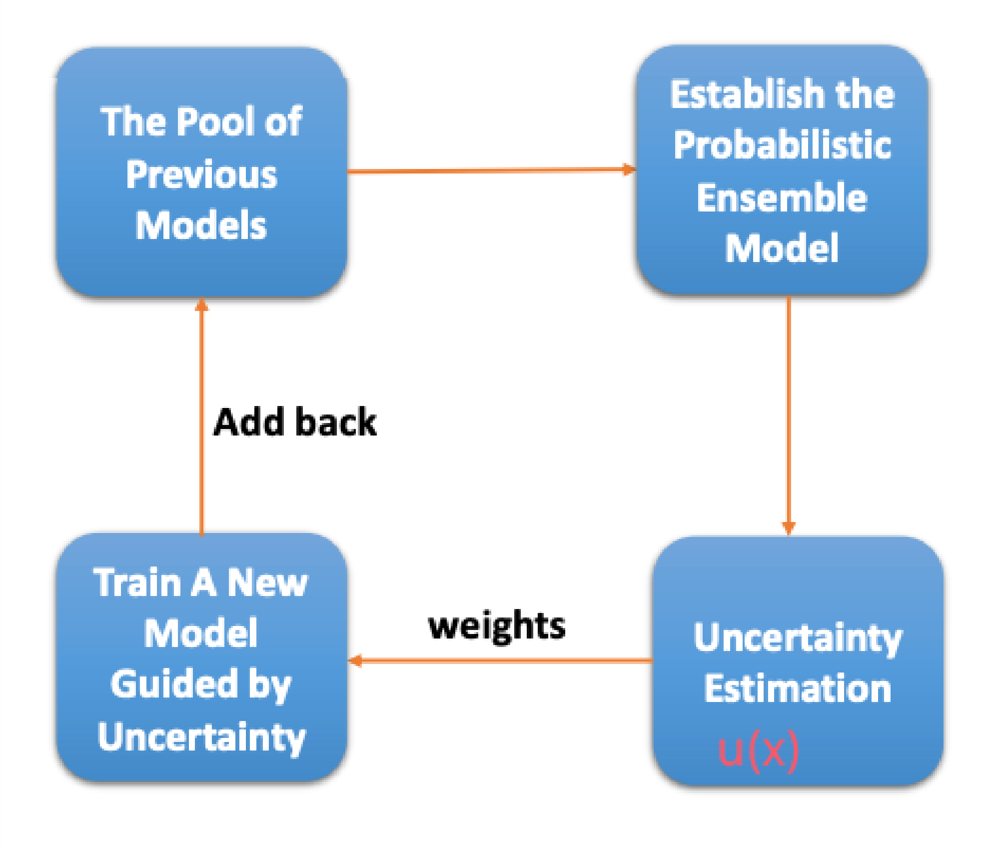
$$p_{PE} = \sum_{i=1}^{N} \lambda_i N(\theta; \theta_i, \Sigma_i) \sim (\mu_P, \Sigma_P)$$

$$\mu_D = \mu_P \quad \Sigma_D < \Sigma_P$$

Proposition 3.5: Overconfidence Reduction of PE

$$\lim_{\eta \to \infty} p_{PE}(y = c | \eta x) \leq \sum_{i=1}^{N} \frac{\lambda_i}{1 + \sum_{j \neq c} \exp\{-t_i^{(j)} - t_i^{(c)}\}} \tag{10}$$

2.3 Adaptive Uncertainty-Guided Ensemble Learning (AUEL)



The uncertainty-guided training loss:

$$L_{nll}(\theta) = -\frac{1}{B}\sum_{m=1}^{B} w(x_m)\log(y_m \mid x_m, \theta),$$

where

$$w(x_m) = \frac{\exp(a * \log(u(x_m)) + b)}{\sum_j^B \exp\big(a * \log\big(u\big(x_j\big)\big) + b\big)}.$$

While a standard Negative Log-likelihood loss is:

$$L_{nll} = \frac{1}{N}\sum_{i=1}^{N} \log(y_i \mid x_i, \theta).$$

2.3 Adaptive Uncertainty-Guided Ensemble Learning (AUEL)

Proposition 3.6: Prediction Error Bound

- The prediction error of the ensemble is bounded by the total uncertainty, providing a theoretical basis for the uncertainty-guided training approach.

Proposition 3.7: Balance with Uncertainty

- For imbalanced classification problems, the model tends to focus on minority classes, ensuring that epistemic uncertainty plays a key role in preventing overconfidence in majority class predictions.

2.4 Mixture of Gaussian Refinement

Parameters waiting tuned: $\left\{\{\lambda_i\}_{i=1}^N, \{\theta_i\}_{i=1}^N, \{\Sigma_i\}_{i=1}^N\right\}$

E-step: construct the loss function $Q(\phi|\phi^0, \mathcal{D})$ as the expected value of the log-likelihood function of ϕ with respect to the current conditional distribution of Z given ϕ^0 and $\mathcal{D}$.

$$\begin{aligned}
\log p(\mathcal{D}|\phi) &= \sum_{m=1}^{M} \log p(\mathcal{D}_m|\phi) \\
&= \sum_{m=1}^{M} \log \sum_{i=1}^{N} \frac{p(Z=i|\mathcal{D}_m, \phi^0)}{p(Z=i|\mathcal{D}_m, \phi^0)} p(\mathcal{D}_m, Z=i|\phi) \\
&\geq \sum_{m=1}^{M} \sum_{i=1}^{N} p(Z=i|\mathcal{D}_m, \phi^0) \log \frac{p(\mathcal{D}_m, Z=i|\phi)}{p(Z=i|\mathcal{D}_m, \phi^0)} \\
&:= Q(\phi|\phi^0, \mathcal{D})
\end{aligned} \tag{13}$$

M-step: maximize $Q(\phi|\phi^0, \mathcal{D})$ with respect to ϕ.

$$\phi^* = \arg\max_{\phi} Q(\phi|\phi^0, \mathcal{D}) \tag{14}$$

2.4 Mixture of Gaussian Refinement

Closed-form solution for $\{\lambda_i^*\}_{i=1}^N$:

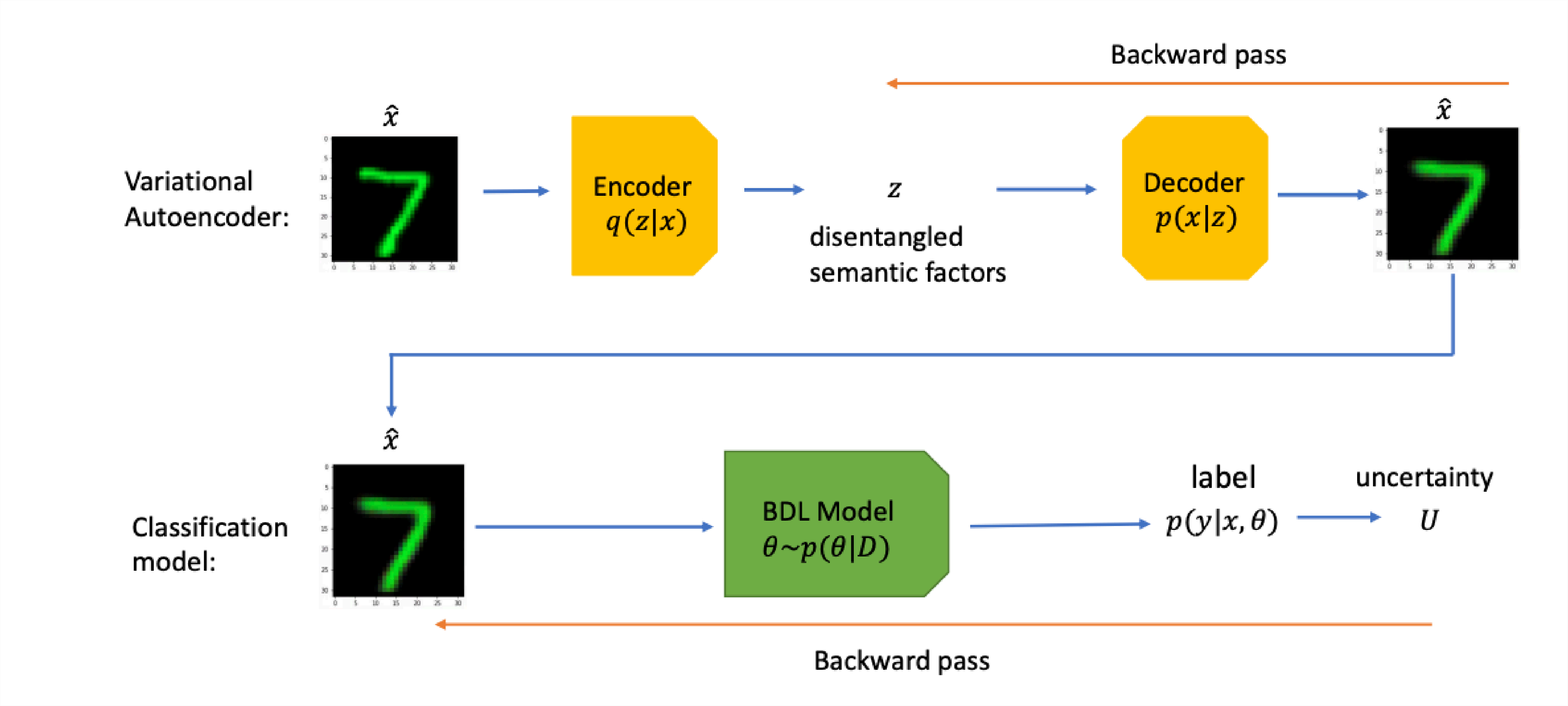
$$\lambda_i^* = \frac{\sum_{m=1}^M p(Z = i|\mathcal{D}_m, \phi^0)}{\sum_{m=1}^M \sum_{j=1}^N p(Z = j|\mathcal{D}_m, \phi^0)} \tag{15}$$

Letting $p_m(\theta) = p(y_m|x_m, \theta)$,

$$p(Z = i|\mathcal{D}_m, \phi^0) = \frac{\lambda_i^0 \int p_m(\theta)\mathcal{N}(\theta; \theta_i^0, \Sigma_i^0)d\theta}{\sum_{j=1}^N \lambda_j^0 \int p_m(\theta)\mathcal{N}(\theta; \theta_j^0, \Sigma_j^0)d\theta} \tag{16}$$

Then given $Z \sim Cat(\{\lambda_i\})$, we assign each data samples to its top l nearest components based on their weighted log-likelihood (i.e.,$l = \frac{N}{2}$).

3. Paper Reading: Semantic Attribution for Explainable UQ


Backward pass
$\hat{x}$
Variational
Autoencoder:
$\hat{x}$
Encoder
$q(z|x)$
z
disentangled
semantic factors
Decoder
$p(x|z)$
$\hat{x}$
Classification
model:
BDL Model
$\theta \sim p(\theta|D)$
label
$p(y|x,\theta)$
uncertainty
U
Backward pass

4. Plans for Next Week

4. Plans for Next Week

1. Hands-on Coding: build basic Resnet/WideResnet/Transformer and do Uncertainty Quantification & Evaluation on them using MC-DropOut and Deep Ensemble. (read original papers before coding)
2. Other paper reading (tentative) plan about Hanjing's work:
   - Uncertainty-Guided Probabilistic Transformer for Complex Action Recognition
   - Beyond Dirichlet-based Models: When Bayesian Neural Networks Meet Evidential Deep Learning
3. A long-term thing: Build up my knowledge in Causal Inference/Discovery (I will talk it with Naiyu later for a tentative study plan.)