Knowledge Distillation via Constrained Variational Inference

Bridging Discriminative Power and Interpretability

Ardavan Saeedi[1], Yuria Utsumi[2], Li Sun[3], Kayhan Batmanghelich[3], Li-wei H. Lehman[2]

[1]Hyperfine [2]MIT [3]University of Pittsburgh

AAAI 2022

Roadmap

The Interpretability-Performance Trade-off

Discriminative Models

- High predictive performance
- Black-box nature
- Limited interpretability
- Dense, complex representations

Probabilistic Graphical Models

- Interpretable structure
- Principled uncertainty quantification
- Poor predictive performance
- Two-stage feature extraction

Core Challenge

Can we distill the discriminative power of neural networks into interpretable graphical models while preserving their generative capabilities?

Knowledge Distillation: Beyond Compression

Teacher Model
(Discriminative)

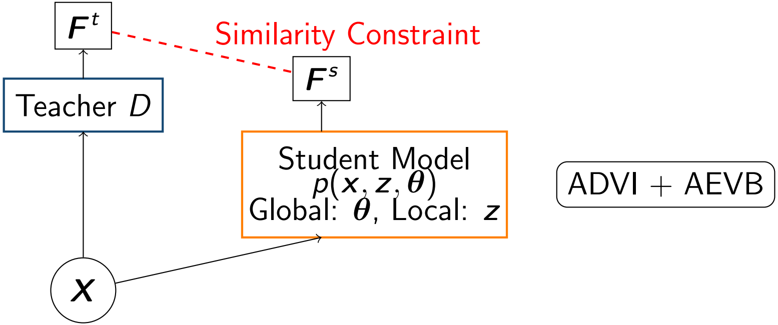
$\downarrow$ Knowledge Distillation

Student Model
(Graphical Model)

Desired Properties

- Interpretability
- Generative modeling
- Predictive power
- Uncertainty quantification

Framework Overview



Key Innovation

Similarity-preserving knowledge distillation constraint integrated into black-box variational inference

Mathematical Framework

Constrained Variational Objective

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\log p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}) - \log q_\phi(\boldsymbol{z}, \boldsymbol{\theta})] \quad (1)$$
$$+ \gamma_\eta \cdot \frac{1}{N^2} \|\bar{\boldsymbol{F}}^s - \bar{\boldsymbol{F}}^t\|_F^2 \quad (2)$$

Similarity Matrices

$$\tilde{\boldsymbol{F}}^s = \boldsymbol{F}^s \cdot (\boldsymbol{F}^s)^T \quad (3)$$
$$\tilde{\boldsymbol{F}}^t = \boldsymbol{F}^t \cdot (\boldsymbol{F}^t)^T \quad (4)$$

$\bar{\boldsymbol{F}}^s, \bar{\boldsymbol{F}}^t$: ℓ_2-normalized versions

Variational Approximation

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\exp(\boldsymbol{\omega})^2)) \quad (5)$$
$$q(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_{\phi_z}(\boldsymbol{x}), \mathrm{diag}(\exp(\boldsymbol{\omega}_{\phi_z}(\boldsymbol{x}))^2)) \quad (6)$$

Black-Box Variational Inference Integration

ADVI for Global Variables θ

- Transform to unconstrained space: $\boldsymbol{\Theta} = T(\theta)$
- Factorized Gaussian approximation
- Automatic differentiation for gradients

AEVB for Local Variables z

- Recognition network: $x \mapsto q(z|x)$
- Amortized inference
- Reparameterization trick

Advantages

- Model-agnostic
- Scalable inference
- Easy implementation
- Broad applicability

Technical Detail

Combined ADVI + AEVB enables handling both global and local latent variables in a unified framework

Application 1: COPD Disease Subtyping

Dataset & Task

- COPDGene: 7,292 subjects
- CT lung images
- Predict clinical severity (FEV1pp, FEV1/FVC, etc.)
- Identify interpretable disease subtypes

Models

- **Teacher**: Subject2Vec (deep sets)
- **Student**: Gaussian-LDA variant
- **Features**: Topic proportions π_s

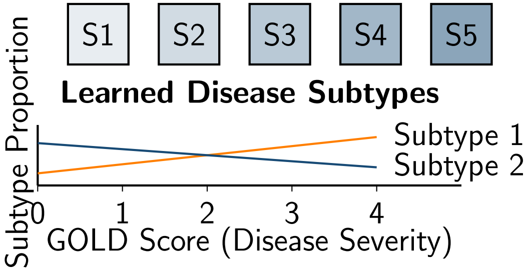
Results: Coefficient of Determination (R^2)

Model	FEV1pp	FEV1/FVC	FVCpp	Distance
G-LDA	0.16	0.30	0.03	0.05
Sup. G-LDA	0.29	0.49	-0.47	0.06
KD-LDA	**0.49**	**0.61**	**0.15**	**0.14**
Teacher	0.65	0.70	0.28	0.16

Key Finding

Substantial improvement in predictive performance while maintaining generative capabilities (ELBO preserved)

COPD Results: Clinical Interpretability



Clinical Insights

- Subtype 1 proportion increases with disease severity (GOLD 0→4)
- Subtype 2 decreases with severity
- Learned subtypes correlate with anatomical regions
- Interpretable disease progression patterns

Application 2: Sepsis Dynamics Modeling

Dataset & Task

- MIMIC-III: 11,648 sepsis patients
- 29 physiological variables
- Predict in-hospital mortality
- Model clinical state transitions

Models

- **Teacher**: LSTM
- **Student**: AR-HMM
- **Features**: State marginals

Results

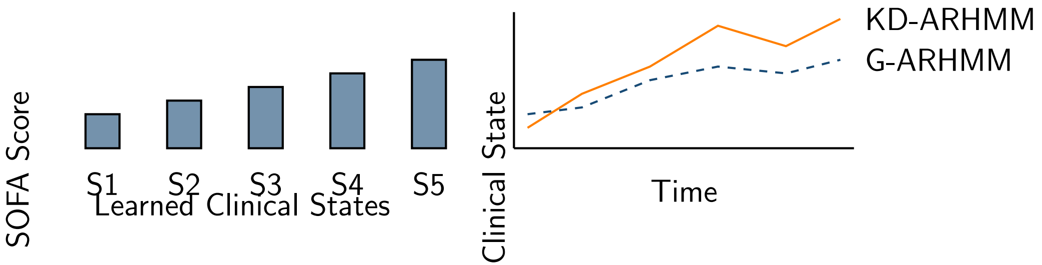
Model	AUROC	ELBO
G-ARHMM	0.56	-55.24
Sup. G-ARHMM	0.56	-55.24
KD-ARHMM	**0.65**	**-3.94**
Teacher (LSTM)	0.71	N/A

Remarkable Result

Improved both predictive performance AND generative modeling (ELBO improvement)

Sepsis Results: Clinical State Analysis

SOFA Score Distributions by State



Clinical Validation

- KD-ARHMM states better correlate with SOFA scores
- More clinically meaningful state transitions
- Improved end-organ dysfunction tracking

Key Technical Innovations

1. **Similarity-Preserving Knowledge Distillation**
   - Matches pairwise similarity matrices instead of direct feature matching
   - Handles heterogeneous feature spaces (neural net ¡-¿ graphical model)
2. **Unified ADVI + AEVB Framework**
   - Enables black-box inference for both global and local variables
   - Model-agnostic approach
3. **Constrained Variational Inference**
   - Integrates knowledge distillation as Lagrangian constraint
   - Balances generative modeling and discriminative power

Broader Impact

Framework applicable to any probabilistic graphical model with local latent variables

Algorithmic Complexity & Scalability

Computational Considerations

- **Similarity Matrix**: $O(N^2)$ space and computation per batch
- **Mini-batch Strategy**: Reduces to $O(B^2)$ where $B \ll N$
- **Gradient Computation**: Automatic differentiation through constraint

Hyperparameter Sensitivity

- **Regularization Weight** γ_η: Controls knowledge distillation strength
- **Batch Size** B: Affects similarity matrix approximation quality
- **Architecture**: Recognition network capacity impacts local variable inference

Convergence Properties

Inherits convergence guarantees from ADVI with additional constraint satisfaction

Limitations & Future Directions

Current Limitations

- BBVI weaknesses:
  - Posterior variance underestimation
  - Initialization sensitivity
  - Amortization gap
- Quadratic scaling in similarity computation
- Hyperparameter tuning complexity

Future Research Directions

- Discrete latent variable support
- Alternative similarity measures
- Multi-teacher distillation
- Theoretical analysis of constraint effects
- Hierarchical model extensions

Open Questions

- How does constraint strength affect posterior approximation quality?
- Can we develop adaptive weighting schemes for γ_η?
- What are the theoretical guarantees for knowledge preservation?

Broader Implications

Methodological Impact

- **Interpretable AI**: Bridge between black-box and interpretable models
- **Transfer Learning**: Novel approach to cross-model knowledge transfer
- **Probabilistic ML**: Enhanced inference for structured models

Application Domains

- **Healthcare**: Disease subtyping, patient monitoring, clinical decision support
- **Natural Language**: Topic modeling with neural guidance
- **Computer Vision**: Interpretable scene understanding
- **Time Series**: Dynamics modeling with neural priors

Paradigm Shift

From choosing between interpretability and performance to achieving both through principled knowledge distillation

Summary

Problem Addressed

How to distill discriminative knowledge into interpretable probabilistic models without sacrificing generative capabilities

Solution Approach

Similarity-preserving knowledge distillation integrated into black-box variational inference via constrained optimization

Key Results

- Substantial predictive performance improvements (COPD: 0.16→0.49 R^2, Sepsis: 0.56→0.65 AUROC)
- Preserved or improved generative modeling capabilities
- Clinically meaningful and interpretable learned representations
- General framework applicable to diverse graphical models

Final Throughts

Take-Home Message

Principled knowledge distillation can bridge the interpretability-performance gap in probabilistic machine learning

Questions & Discussion

Thank you for your attention

av.saeedi@gmail.com