Co-Training for Policy Learning

Jialin Song
Ravi Lanka
Yisong Yue
Hiro Ono
Policy Learning (Reinforcement & Imitation)

**Goal:** Find “Optimal” Policy

**Imitation Learning:**
Optimize imitation loss

**Reinforcement Learning:**
Optimize environmental reward

Diagram:
- Agent
  - State/Context $s_t$ → Action $a_t$
  - $s_{t+1}$ from Environment / World
Policy Learning is Hard

- Long time horizons
- Sparse or expensive feedback
- Exponential in time horizon
  - (reinforcement learning)
- Infeasible to obtain sufficient demonstrations
  - (imitation learning)
Example: Learning to Search (Combinatorial Optimization)

\[
\max - \sum_{i=1}^{5} x_i,
\]

subject to:
\[
\begin{align*}
& x_1 + x_2 \geq 1, \\
& x_2 + x_3 \geq 1, \\
& x_3 + x_4 \geq 1, \\
& x_3 + x_5 \geq 1, \\
& x_4 + x_5 \geq 1, \\
& x_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}
\end{align*}
\]

[He et al., 2014] [Song et al., arXiv]
Learning from Multiple Views

Example: Minimum Vertex Cover

\[
\begin{align*}
\text{max} & \quad - \sum_{i=1}^{5} x_i, \\
\text{subject to:} & \\
x_1 + x_2 & \geq 1, \\
x_2 + x_3 & \geq 1, \\
x_3 + x_4 & \geq 1, \\
x_3 + x_5 & \geq 1, \\
x_4 + x_5 & \geq 1, \\
x_i & \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}
\end{align*}
\]

Graph View

[He et al., 2014]

Integer Program View (Branch & Bound View)

[Khalil et al., 2017]
Learning from Multiple Views

Example: Different Types of Integer Programs

ILP

QCQP
Co-Training [Blum & Mitchell, 1998]

- Many learning problems have different sources of information
- Webpage Classification: Words vs Hyperlinks

(Taken from Avrim Blum’s slides)
Semi-Supervised Regression with Co-Training

Zhi-Hua Zhou and Ming Li
National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{zhouzh, llim}@lamda.nju.educn

A New Analysis of Co-Training

Wee Wang
Zhi-Hua Z
National K

Applying Co-Training methods to Statistical Parsing

Anoop Sarkar
Dept. of Computer and Information Science
University of Pennsylvania
200 South 33rd Street,
Philadelphia, PA 19104-6389 USA
appliong@cs.upenn.edu

Bayesian Co-Training

Shipeng Yu
Balaji Krishnapuram
Business Intelligence and Analytics
Siemens Medical Solutions USA, Inc.
51 Valley Stream Parkway
Malvern, PA 19355, USA

Rome Rosales
BOMER@YAHOO.COM

A Co-training Approach for Multi-view Spectral Clustering

Abhishek Kumar
Hal Daumee III
Department of Computer Science, University of Maryland, College Park, MD 20742, USA

Co-Training for Domain Adaptation

Minmin Chen, Kilian Q. Weinberger
Department of Computer Science and Engineering
Washington University in St. Louis
St. Louis, MO 63130
mcl5,kilian@wustl.edu

Co-Training and Expansion: Towards Bridging Theory and Practice

Understanding the Behavior of Co-training

Svetlana Kiritchenko and Stan Matwin
School of Information Technology and Engineering
University of Ottawa
Ottawa, ON, Canada
{svkirs,etan}@site.uottawa.ca

Applying Co-Training to Reference Resolution

María-Florencia C.
Computer Scienc
Carnegie Meln

Applying Co-Training to Reference Resolution

Risto Müller
Villa Bosch
Waldstrassenweg 33
heidelberg, Germany
ml.villa-bosch.de

PAC Generalization Bounds for Co-training

Sanjey Dasgupta
AT&T Labs-Research
dasgupta@research.att.com

Michael L. Littman
AT&T Labs-Research
mlittman@research.att.com

David McAllester
AT&T Labs-Research
dmace@research.att.com

Unsupervised Improvement of Visual Detectors using Co-Training

Michael Strube
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
mstrube@cs.cmu.edu

Reinforced Co-Training

William Yang Wang
Department of Computer Science
University of California
Santa Barbara, CA 93106 USA
william@cs.ucsb.edu

Email Classification with Co-Training

Rayid Ghani
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
rayid@cs.cmu.edu

Kamal Nigam

Paul Viola
Microsoft Research
One Microsoft Way
Redmond, WA 98052
violamicrosoft.com

Yuval Freund
Computer Science Dept.
Columbia University
New York, NY 10027
freund@columbia.edu

Anat Levin'
School of CS and Eng.
The Hebrew University
91904 Jerusalem, Israel
alevin@cs.huji.ac.il

Co-training for Policy Learning

Jialin Song† Ravi Lanka† Yisong Yue† Masahiro Ono†
†California Institute of Technology
‡Jet Propulsion Laboratory, California Institute of Technology

A Co-training Approach for Multi-view Spectral Clustering

PAC Generalization Bounds for Co-training

Shi-Yu Sun
ECE, University of Texas at Dallas

Emmanouil Vlachos
Department of Computer Science, University of Maryland, College Park, MD 20742, USA
What’s Different about Policy Co-Training?

• Sequential Decisions vs 1-Shot Decisions

• (Sparse) Environmental Feedback
  • Can collect more “labels”

• Different Action Spaces
  • Graph vs Branch-and-Bound

\[
\begin{align*}
\max & - \sum_{i=1}^{5} x_i, \\
\text{subject to:} & \quad x_1 + x_2 \geq 1, \\
& \quad x_2 + x_3 \geq 1, \\
& \quad x_3 + x_4 \geq 1, \\
& \quad x_3 + x_5 \geq 1, \\
& \quad x_4 + x_5 \geq 1, \\
& \quad x_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}
\end{align*}
\]
Intuition

E.g., [1]

MVC Instance

E.g., [2,3]

[1] “Learning combinatorial optimization algorithms over graphs” [Khalil et al., 2017]
Intuition

MVC Instance

E.g., [1]

E.g., [2,3]

[1] “Learning combinatorial optimization algorithms over graphs” [Khalil et al., 2017]
Intuition

E.g., [1]

MVC Instance

E.g., [2,3]

[1] “Learning combinatorial optimization algorithms over graphs” [Khalil et al., 2017]
Theoretical Insight

• Different representations differ in hardness
• Goal: quantify improvement

\[ \Omega : \text{all problems} \]

\[ \Omega_1 : \text{representation 1 easier} \]

\[ \Omega_2 : \text{representation 2 easier} \]
(Towards) a Theory of Policy Co-Training

• Two MDP “views”: $M^1$ & $M^2$
  • $f^{1\to2}(\tau^1) \Rightarrow \tau^2$ (and vice versa)
  • “Trajectory” / “Rollout”
  • Realizing $\tau^1$ on $M^1 \iff$ realizing $\tau^2$ on $M^2$

• **Question:** when does having two views/policies help?
  • Policy Improvement (next slide)
    • Builds upon [Kang et al., ICML 2018]
  • Optimality Gap for Shared Action Spaces (in paper)
    • Builds upon [DasGupta et al., NeurIPS 2002]
Policy Improvement Bound

Standard for Policy Gradient

$J(\pi'^1) \geq J_{\pi^1}(\pi'^1) - \frac{2\gamma(\alpha^1_\Omega \varepsilon^1_\Omega + 4\beta^2_\Omega_2 \varepsilon^2_\Omega_2)}{(1 - \gamma)^2} + \delta^2_{\Omega_2}$

1-step suboptimality of $\pi^1$ on $\Omega$

KL Divergence of $\pi^1$ vs $\pi'^1$ on $\Omega$

Performance of new policy (either RL or IL)

Discount

Performance Gap of $\pi^2$ over $\pi^1$ on $\Omega_2$: $J(\pi^2 | M \sim \Omega_2) - J(\pi^1 | M \sim \Omega_2)$

JS Divergence of $\pi^2$ vs $\pi^1$ on $\Omega_2$

1-step suboptimality of $\pi^2$ on $\Omega_2$

Want to Minimize

Want to Maximize

Builds upon theoretical results from [Kang et al., ICML 2018]
Policy Improvement Bound (Summary)

\[ J(\pi'_{1}) \geq J_{\pi^{1}}(\pi'_{1}) - \frac{2\gamma (\alpha_{\Omega}^{1} \epsilon_{\Omega}^{1} + 4\beta_{\Omega_2}^{2} \epsilon_{\Omega_2}^{2})}{(1 - \gamma)^{2}} + \delta_{\Omega_2}^{2} \]

- Minimizing \( \beta_{\Omega_2}^{2} \) → low disagreement between \( \pi^{2} \) vs \( \pi^{1} \)

- Maximizing \( \delta_{\Omega_2}^{2} \) → high performance gap \( \pi^{2} \) over \( \pi^{1} \) on some MDPs
CoPiEr Algorithm (Co-training for Policy Learning)

**Update** (only showing 1 view)

Augmented Obj: \( \tilde{J}(\pi') = J_\pi(\pi') - \lambda L(\pi', \tau') \)

Take gradient step

**Exchange** (only showing 1 version)

If \( \pi^1 \) better: \( \tau'^2 = f^{1\to2}(\tau^1), \tau'^1 = \emptyset \)

If \( \pi^2 \) better: \( \tau'^1 = f^{2\to1}(\tau^2), \tau'^2 = \emptyset \)

**Rollout**

Run \( \pi^1 \to \tau^1 \)

Run \( \pi^2 \to \tau^2 \)

**Augmented Obj:**

\[
J_{F}(\pi, G) = J(\pi) - \lambda L(\pi, \tau) - \mu H(\pi, \tau)
\]

```
max - \sum_{i=1}^{5} \delta_i,
subject to:
\delta_1 + \delta_2 \geq 1,
\delta_2 + \delta_3 \geq 1,
\delta_3 + \delta_4 \geq 1,
\delta_3 + \delta_5 \geq 1,
\delta_4 + \delta_5 \geq 1,
\delta_i \in \{0, 1\}, \forall i \in \{1, \cdots, 5\}
```
Performance comparison for Minimum Vertex Cover

Strong vs Baselines (w/o Co-Training)

CoPiEr Final Outperforms Individual Views

Strong vs Gurobi

Erdős–Rényi (100-500 vertices)

RL on Graph View
[He et al., 2014]

IL on MILP View
[Khalil et al., 2017]

More experiments in paper
Co-Training for Policy Learning (summary)

• First formal framework for policy co-training

• Novel theoretical insights

• Principled algorithm design

• Strong experimental results

\[ J(\pi'(1)) \geq J_{\pi'(1)} - \frac{2\gamma(\alpha_1^1 e_1^1 + 4\beta_2^2 e_2^2)}{(1 - \gamma)^2} + \delta_{12}^2 \]
References

• “Co-Training for Policy Learning,” Jialin Song, Ravi Lanka, Yisong Yue, Masahiro Ono, UAI 2019

• “Combining Labeled and Unlabeled data with Co-training,” Avrim Blum, Tom Mitchell, COLT 1998

• “PAC Generalization Bounds for Co-training,” Sanjoy DasGupta, Michael Littman, David McAllester, NeurIPS 2002

• “Policy Optimization with Demonstrations,” Bingyi Kang, Zequn Jie, Jiashi Feng, ICML 2018

• “Learning Combinatorial Optimization over Graphs,” Elias Khalil, Hanjun Dai, Yuyu Zhang, Bistra Dilkina, Le Song, NeurIPS 2017

• “Learning to Search in Branch and Bound Algorithms,” He He, Hal Daume III, Jason Eisner, NeurIPS 2014

• “Learning to Search via Retrospective Imitation,” Jialin Song, Ravi Lanka, Albert Zhao, Aadyot Bhatnagar, Yisong Yue, Masahiro Ono, arXiv
Extra Slides
Notation

• MDP: $M = (S, A, P, r, \gamma, S_T)$

• Policy (or agent): $\pi(s) \rightarrow a$
  • $V_\pi(s) = E_\tau[\sum_{t=0}^{\infty} \gamma^tr(s_t, a_t)|s_0 = s]$
  • $Q_\pi(s, a) = E_\tau[\sum_{t=0}^{\infty} \gamma^tr(s_t, a_t)|s_0 = s, a_0 = a]$
  • $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$

• Goal: maximize $J(\pi) \equiv E_{M,s_0}[V_\pi(s_0)]$

(Over Distribution of MDPs)
Policy Improvement Bound (detailed)

\[ J(\pi'') \geq J(\pi') - \frac{2\gamma(4\beta^2 \Omega_2 \varepsilon^2 \Omega_2 + \alpha^1 \Omega \varepsilon^1 \Omega)}{(1 - \gamma)^2} + \delta^2_{\Omega_2} \]

New policy
Sampled via \( \pi^1 \)
Discount

Builds upon theoretical results from [Kang et al., ICML 2018]
Performance comparison for Risk-Aware Path Planning

- **Chance-Constrained Path Planning**
  - [Ono & Williams, 2008]
- **IL on MILP**
- **IL on QCQP**
  - [He et al., 2014]
- **IL on QCQP**
  - [Song et al., arXiv]
OpenAI Gym & Mujoco

Partitioned state space into two views

Shared action space

RL on both views