Improving Multi-Task Generalization via Regularizing Spurious Correlation

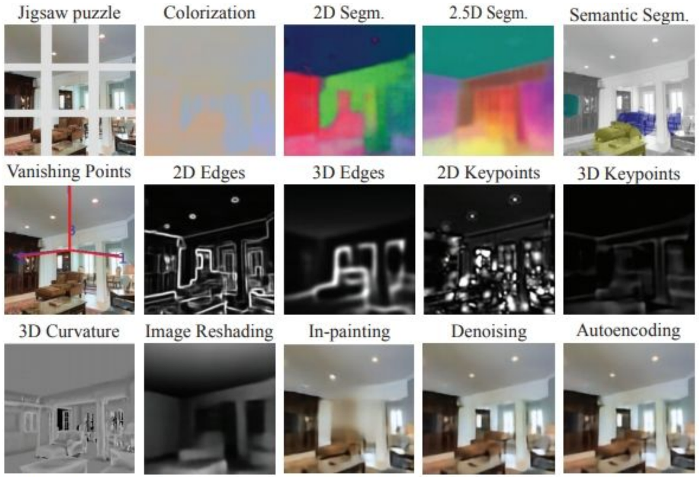
UCLA

Google Research

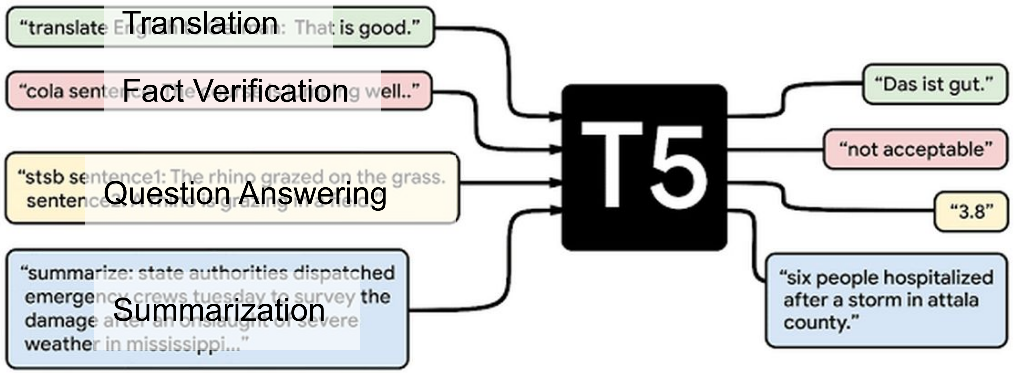
Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, Ed H. Chi

Do Multi-Task Learning **always** benefits Generalization?

- Multi-Task Representation Learning aims at training a neural network encoders that could get **representations** that are informative to handle multiple tasks simultaneously.



Taskonomy: Disentangling Task Transfer Learning, CVPR 2018



Google **T5** (Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, JMLR 2020)

Do Multi-Task Learning **always** benefits Generalization?

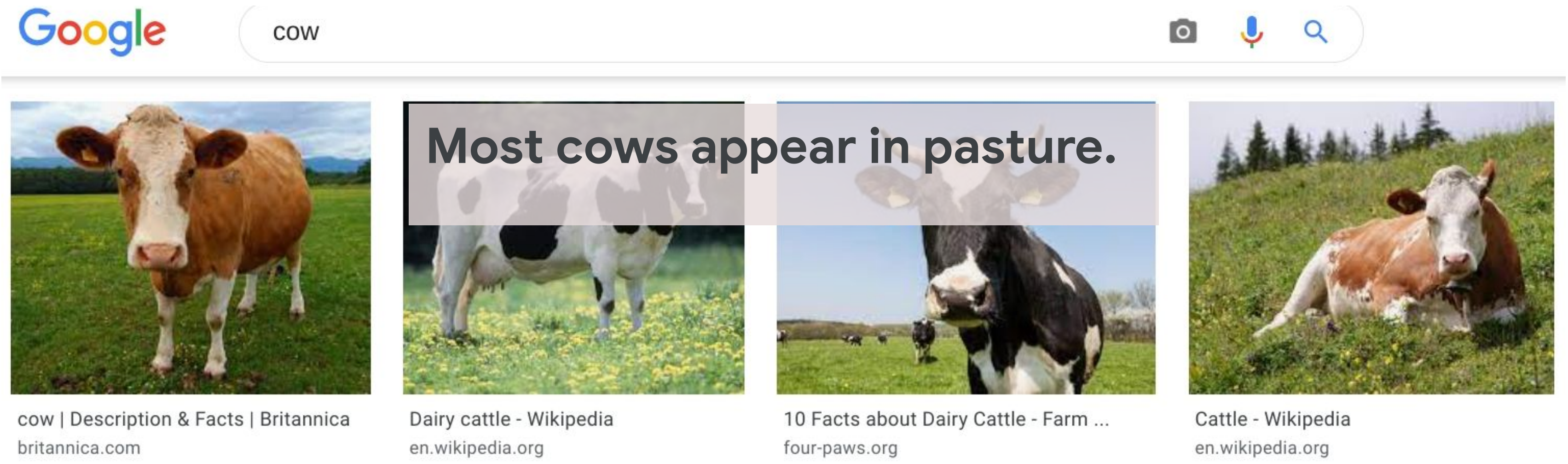
- Many empirical results [1,2] show that there exist **negative transfer** when we train two tasks together, even if the two tasks are semantically correlated.

Trained With	Relative Performance On: SemSeg	Depth	Normals	Keypoints	Edges	Average
SemSeg	–	3.00%	-2.79%	-5.20%	27.80%	5.70%
Depth	1.72%	–	1.18%	-3.52%	25.73%	6.28%
Normals	10.81%	7.12%	–	88.98%	71.59%	44.62%
Keypoints	3.12%	-0.41%	-10.12%	–	61.07%	13.42%
Edges	0.03%	-1.40%	-4.78%	-3.05%	–	-2.30%
	3.92%	2.08%	-4.13%	19.30%	46.54%	13.54%

- Even with an over-parameterized model that achieves low training error, the final MTL generalization could be even worse than single-task learning.

[1] Which Tasks Should Be Learned Together in Multi-task Learning?", Standley et al. ICML 2020.
[2] A Survey on Negative Transfer, Zhang et al. Trans Neural Netw Learn Syst.

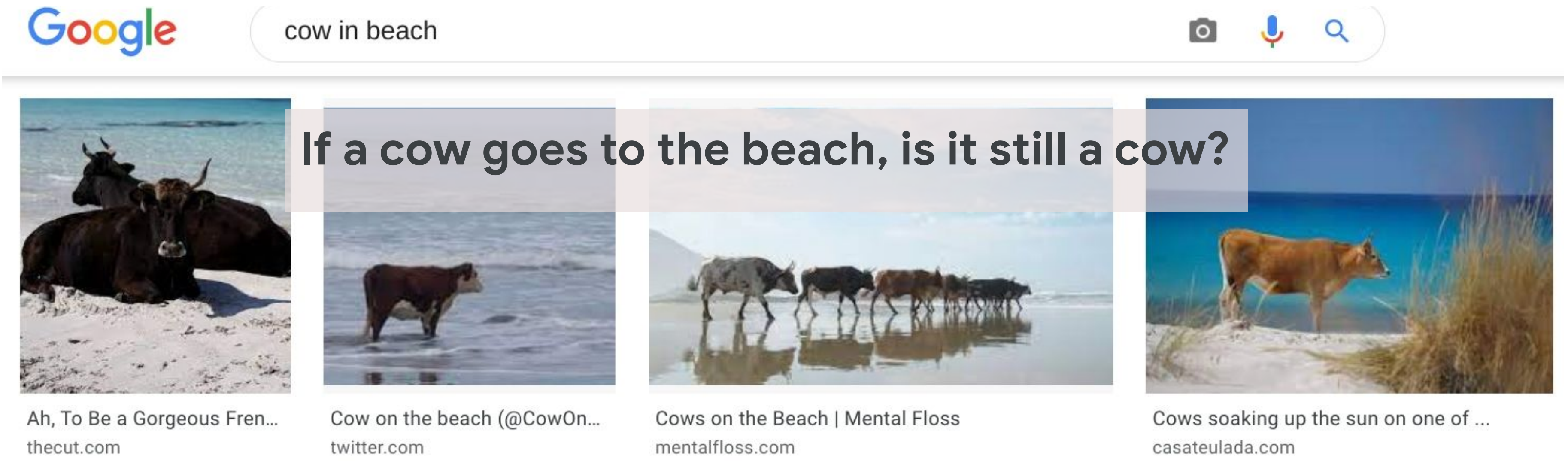
Spurious Correlation Hurts Generalization



- Spurious Features are those non-causal to the target task, but often exists in the training dataset, mostly due to data selection bias.

More examples and discussions about spurious feature could be found in: Geirhos et al. [Shortcut Learning in Deep Neural Networks]

Spurious Correlation Hurts Generalization



- Model is prone to use these feature to fit training data, which hurts generalization [1, 2].
- Two types of spurious feature:
  - independent to task-label (noise);
  - spuriously correlate to label in training set, and the correlation may change in other dataset.

[1] Understanding the Failure Modes of Out-of-Distribution Generalization. Nagarajan et al. ICLR 2021
[2] Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately. Khani et al. FAccT 2021.

Existing Techniques to avoid using spurious features

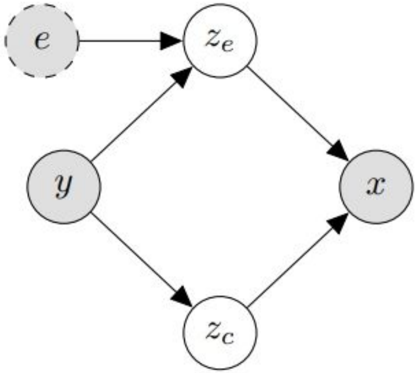


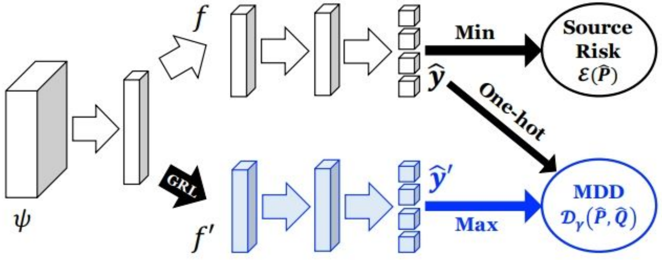
Figure from “The Risks of Invariant Risk Minimization” Elan et al.

- Most existing works only study a single type of spurious feature (e).
- Gender, racial bias, environment, ……

- Adversarial Removal of Spurious Feature in Raw Data Input

[1] Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. Wang et al. ICCV 2019.

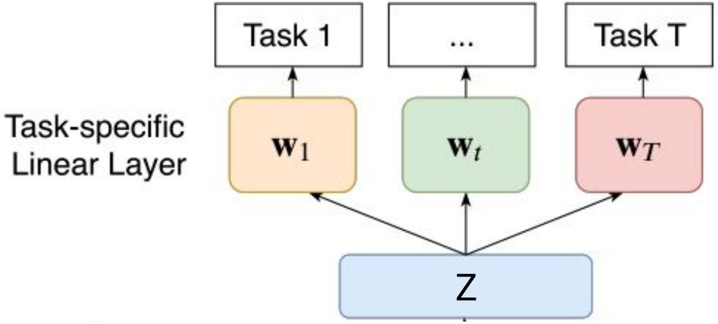
- Learning Domain-Invariant Representation given multiple Domain

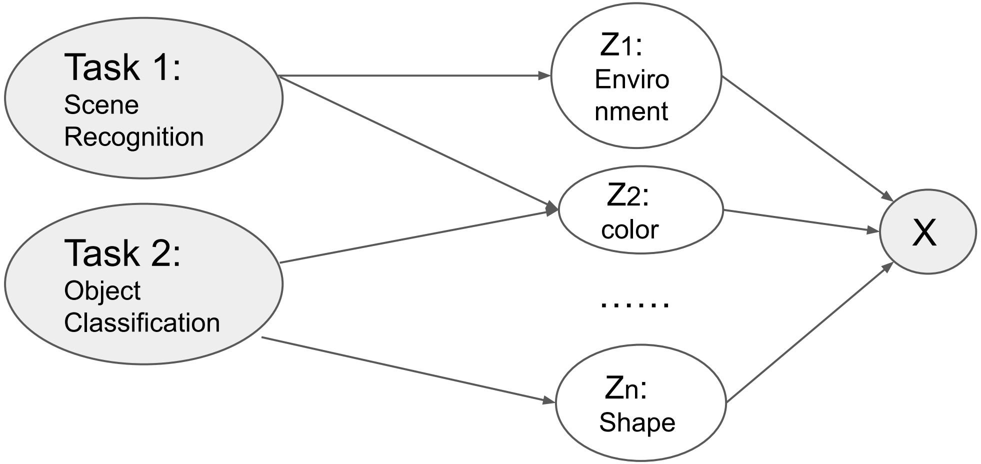


[2] Bridging Theory and Algorithm for Domain Adaptation. Zhang et al. ICML 2019.

Google

Challenges of Spurious correlation in Multi-Task Learning

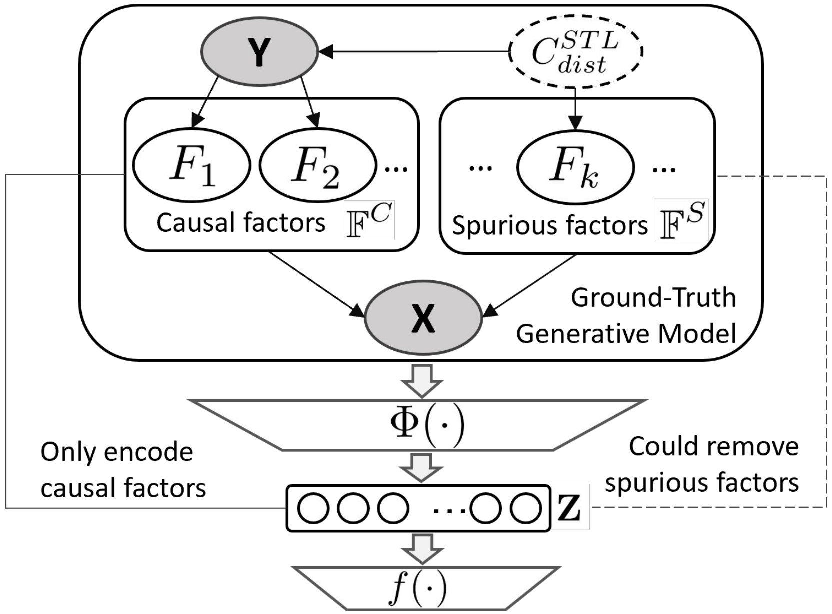




Illustrative Diagram of Causal Generative Model in MTL setting

- the shared MTL model needs to encode all knowledge from different tasks, and **causal** knowledge for one task could be **potentially spurious** to the other.

Spurious Correlation in Single-Task Learning

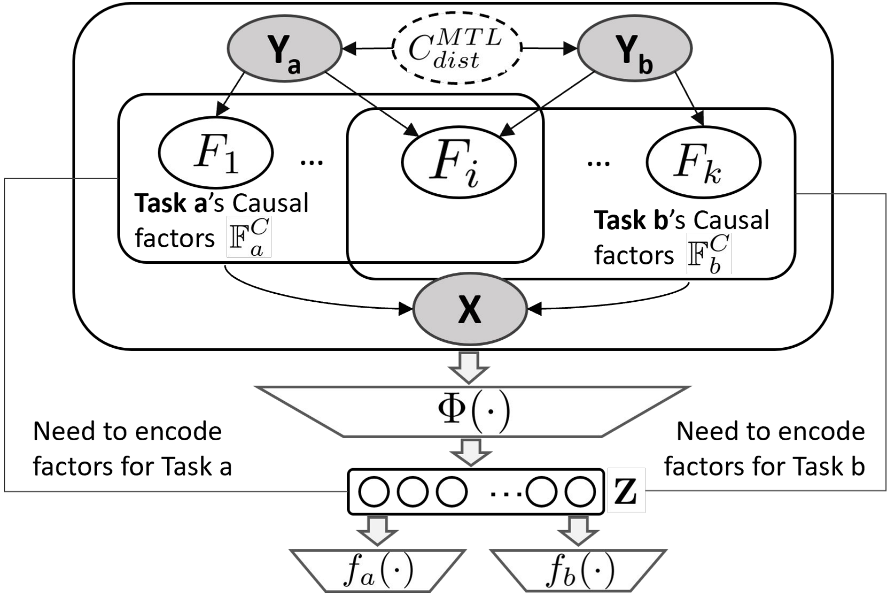


- Spurious correlation in Single-Task Learning is mainly caused by factor-label confounders.
- We could remove spurious factors from representation Z

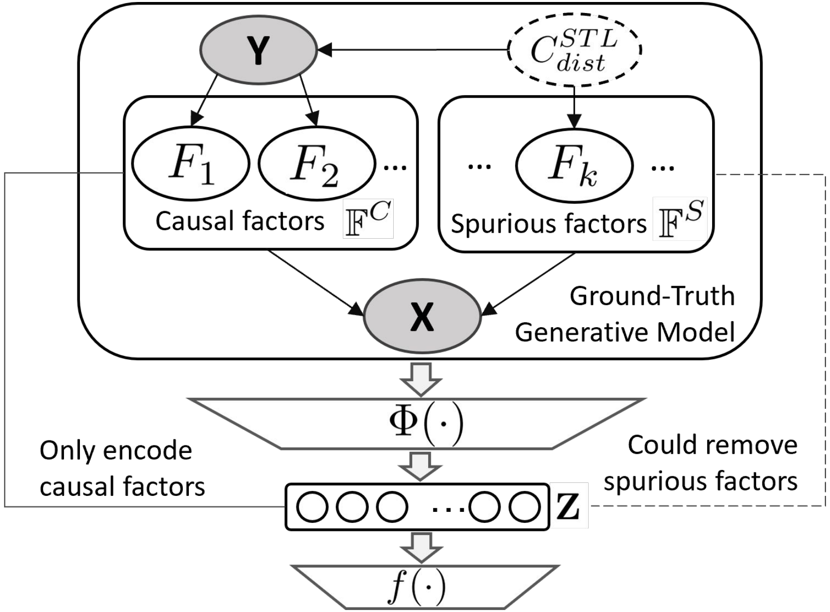
Spurious Correlation in Multi-Task Learning

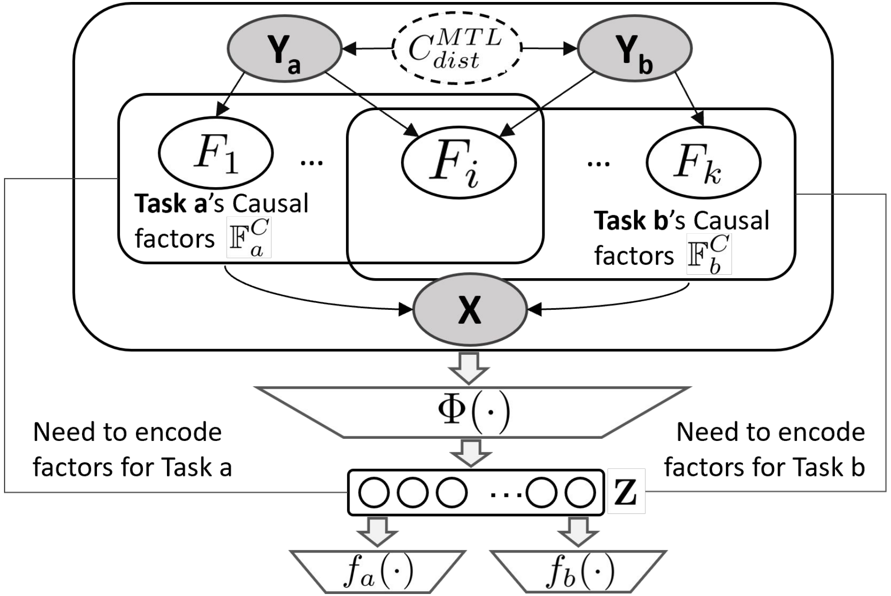
- Spurious correlation in Multi-Task Learning could be caused by label-label confounders.

- Factors for all tasks need to be encoded in share representation, and potentially spurious



Challenges of Spurious correlation in Multi-Task Learning



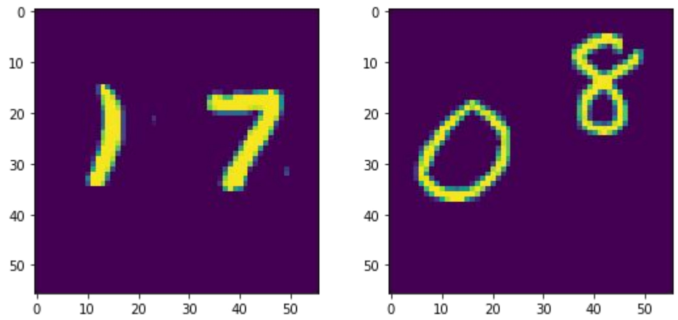


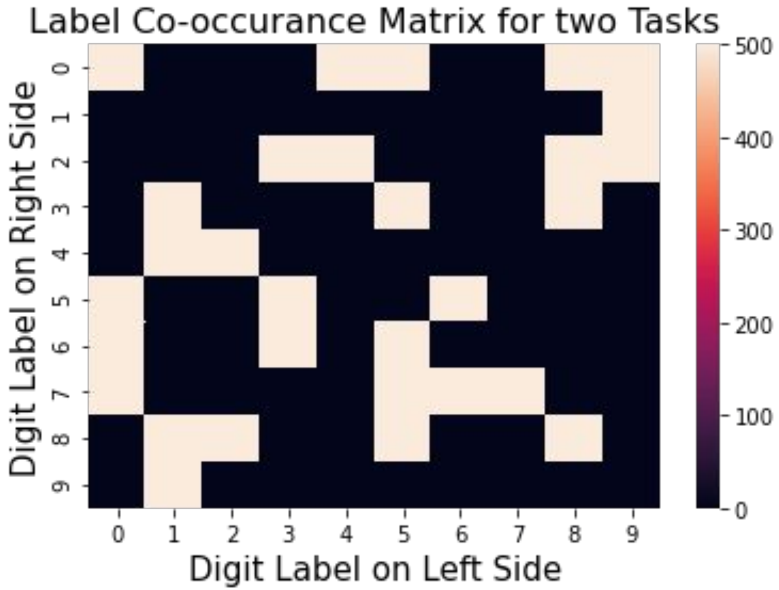
Proposition 1 *Given $m_C \neq 0.5$, the Bayes Optimal per-task classifier has non-zero weights to non-causal factor. Given $m_C = 0.5$ and limited training dataset, the trained per-task classifier will assign non-zero weights to non-causal factor as noise.*

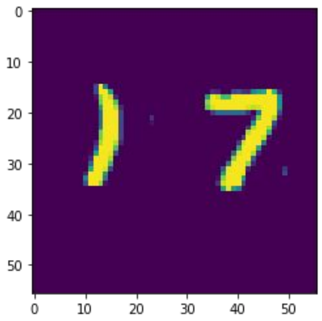
Empirical Analysis to study spurious correlation in MTL

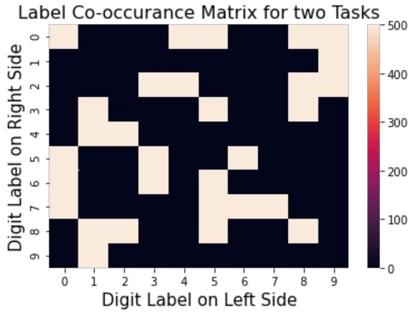
- we use the gradient map to quantify how each task use the feature and spurious ratio

$$Grad(F) = \sum_{(x(\mathbb{F}),y)\in D} \left| \frac{\partial \left(f(\Phi(x))[y] \right)}{\partial F} \right| \qquad \rho_{spur} = \frac{\sum_{F\in \mathbb{F}^S} Grad(F)}{\sum_{F\in \mathbb{F}} Grad(F)}$$



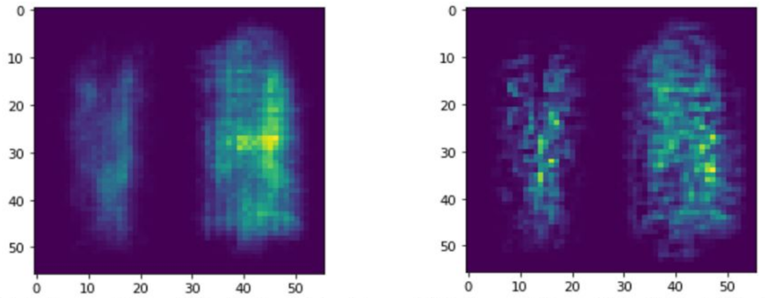






$$Grad(F) = \sum_{(x(\mathbb{F}),y)\in D} \left| \frac{\partial\Big(f(\Phi(x))[y]\Big)}{\partial F} \right|$$

$$\rho_{spur} = \frac{\sum_{F\in\mathbb{F}^S} Grad(F)}{\sum_{F\in\mathbb{F}} Grad(F)}$$



(a) Saliency Map of Single-Task Model (b) Saliency Map of Multi-Task Model

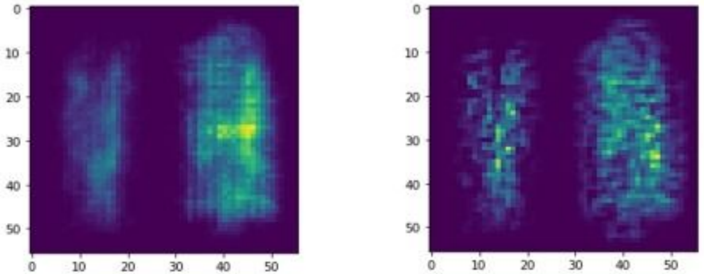
	Multi-SEM		**Multi-MNIST**	
	STL	MTL	STL	MTL
Acc_{train}	0.931	0.936	0.981	0.987
Acc_{val}	0.906	0.882	0.874	0.846
ρ_{spur}	0.128	0.246	0.261	0.328

Empirical Analysis to study spurious correlation in MTL

- we use the gradient map to quantify how each task use the feature and spurious ratio

$$Grad(F) = \sum_{(x(\mathbb{F}),y)\in D} \left| \frac{\partial\left(f(\Phi(x))[y]\right)}{\partial F} \right| \qquad \rho_{spur} = \frac{\sum_{F\in\mathbb{F}^S} Grad(F)}{\sum_{F\in\mathbb{F}} Grad(F)}$$

- By conducting analysis on Multi-MNIST dataset with spurious correlation in training set, we found MTL indeed utilize more spurious feature and influence performance.



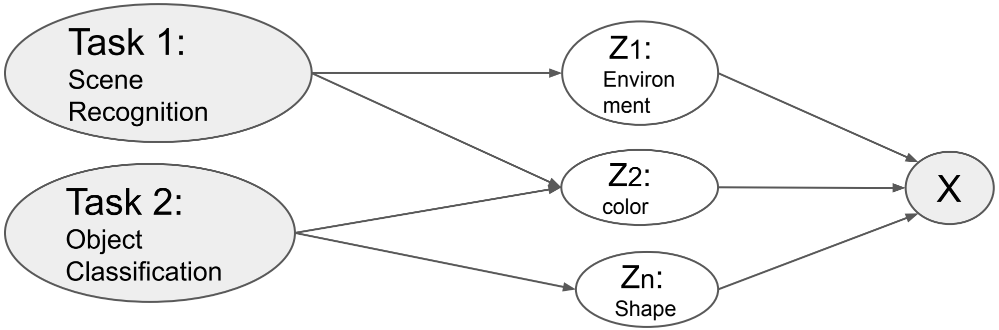
(a) Saliency Map of Single-Task Model (b) Saliency Map of Multi-Task Model

Figure 3: The gradient saliency map of right-side digit classifier. The model trained by MTL exploits left pixels (spurious) more.

	Multi-SEM		**Multi-MNIST**	
	STL	MTL	STL	MTL
Acc_{train}	0.931	0.936	0.981	0.987
Acc_{val}	0.906	0.882	0.874	0.846
ρ_{spur}	0.128	0.246	0.261	0.328

Table 1: Empirical results of multi-task (MTL) and single-task learning (STL) model on synthetic datasets with changing C_{dist}^{MTL}.

Our solution: Multi-Task Causal Representation Learning



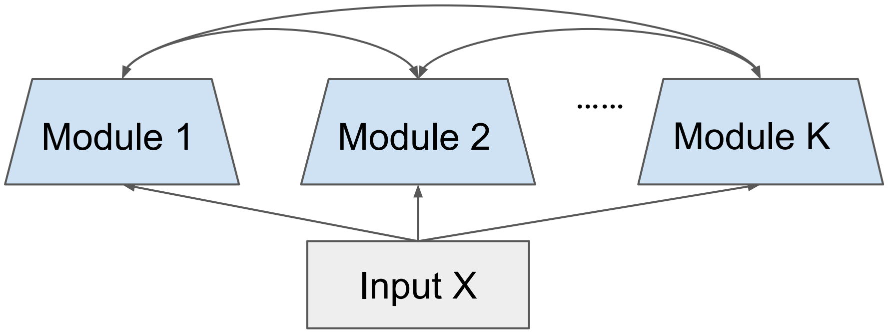
- Motivated by the ground-truth causal generative process, we aim to use a neural model to learn the different data factors and causal relationship between tasks and these factors.

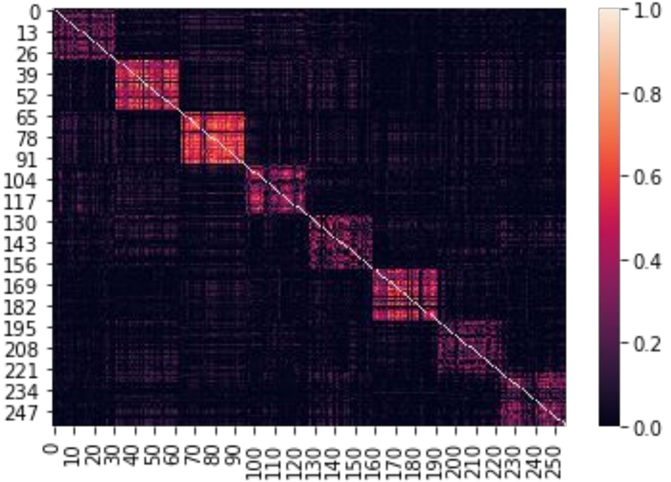
Our solution: Multi-Task Causal Representation Learning

Overall Workflow of MT-CRL:

- Aims to represent multi-task knowledge via **disentangled neural modules**

$$\rho(\mathbf{Z}_i, \mathbf{Z}_j) = \frac{Cov(\mathbf{Z}_i, \mathbf{Z}_j)}{\sqrt{Cov(\mathbf{Z}_i, \mathbf{Z}_i)}\sqrt{Cov(\mathbf{Z}_j, \mathbf{Z}_j)}}$$

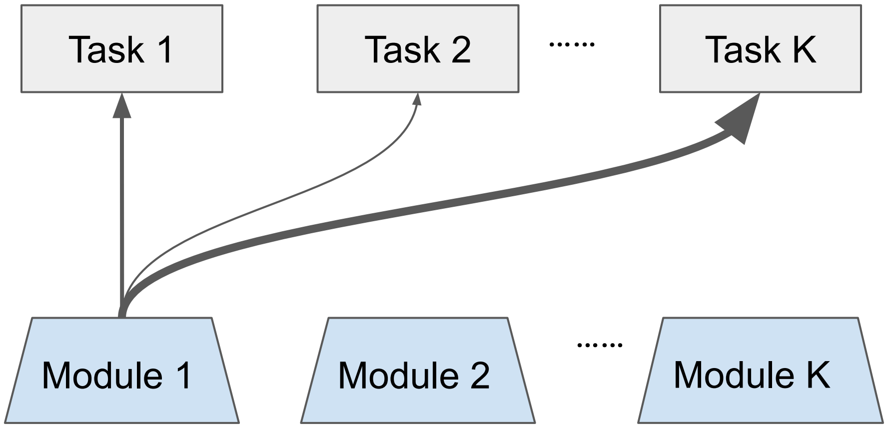




Our solution: Multi-Task Causal Representation Learning

Overall Workflow of MT-CRL:

- Aims to represent multi-task knowledge via **disentangled neural modules**
- Learn robust **task-to-module routing graph** weights via MTL-specific invariant regularization (force graph weights optimal across environments)



task-to-module routing graph regularization:

$$\mathcal{L}_{graph}(A) = \lambda_{sps} \cdot ||A||_1 - \lambda_{bal} \cdot \text{Entropy}\left(\frac{\sum_t A_{t,*}}{\sum_{t,i} A_{t,i}}\right)$$

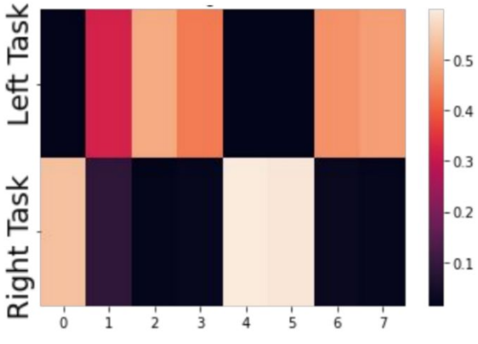
Graph-invariant Risk Minimization (G-IRM)

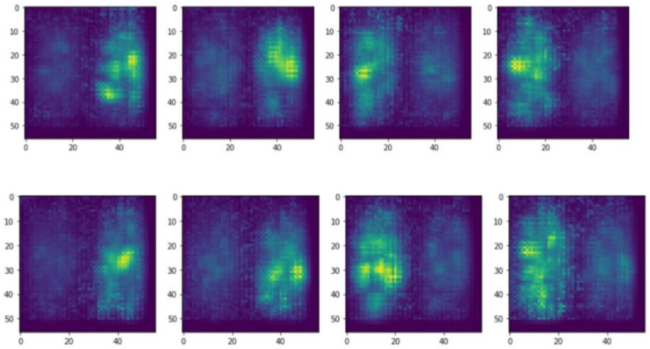
$$\min_{\Phi, A, f} \left(\tilde{\mathcal{L}}(\Phi, A, f) + \lambda_{G\text{-}IRM} \cdot \mathcal{L}_{G\text{-}IRM}(\Phi, A|f)\right)$$

$$\mathcal{L}^{Var}_{G\text{-}IRM}(\Phi, A|f) = \sum_{t \in \mathcal{T}} \sum_{e \in \mathcal{E}} \frac{1}{|\mathcal{E}|} \left\| \nabla_{A=A_t} R^e_t(\Phi, A, f_t) - \text{Avg}_e\left(\nabla_{A=A_t} R^e_t\right) \right\|^2$$

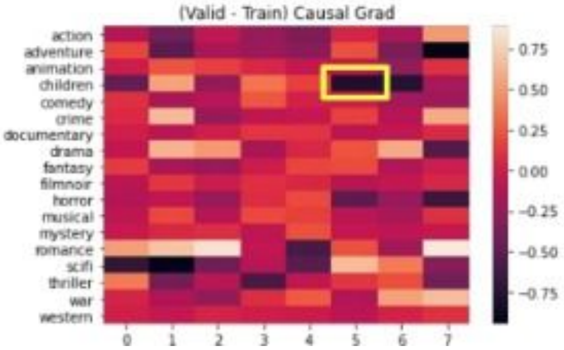
Experiment Results of MT-CRL

Methods	Multi-MNIST	MovieLens	Taskonomy	CityScape	NYUv2	Avg.
Vanilla MTL	(—baseline to calculate relative improvement—)					
Single-Task Learning	+3.3%	+0.2%	-2.5%	-2.4%	-12.2%	-2.7%
MTL + PCGrad	+4.5%	+0.2%	+3.1%	+2.1%	+7.4%	+3.5%
MTL + GradVac	+4.6%	+0.3%	+3.5%	+2.1%	+7.2%	+3.5%
MTL + DANN	+4.1%	+0.4%	+1.2%	+0.3%	-0.4%	+1.1%
MTL + IRM	+5.0%	+0.4%	+1.1%	+0.6%	-0.1%	+1.4%
MT-CRL w/o $\mathcal{L}_{G\text{-}IRM}$	+5.9%	+0.2%	+3.2%	+1.5%	+4.3%	+3.0%
MT-CRL with $\mathcal{L}_{G\text{-}IRM}^{Norm}$	+7.8%	+1.0%	+6.5%	**+2.9%**	+8.0%	+5.2%
MT-CRL with $\mathcal{L}_{G\text{-}IRM}^{Var}$	**+8.1%**	**+1.1%**	**+7.1%**	+2.8%	**+8.2%**	**+5.5%**





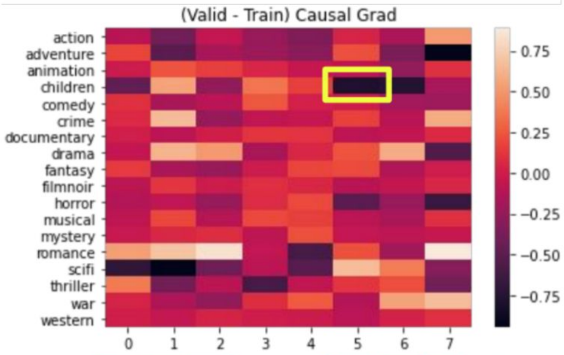
MT-CRL can alleviate spurious correlation



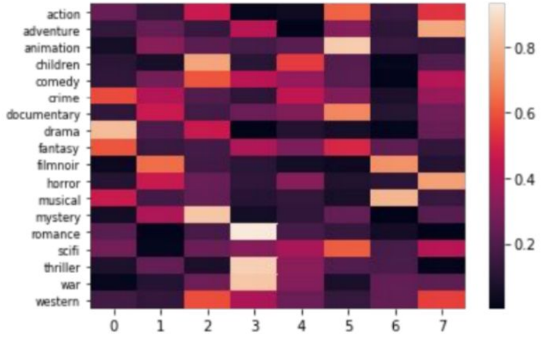
View 1: Shine, go, shawshank, psycho, dumber
View 2: Rocky, october, casino, muppet, payback
View 3: forrest, gump, carrie, now, saving
View 4: i, house, monty, at, life, dark
View 5: good, club, young, stripes, die
View 6: 1978, out, witness, shining, chocolate
View 7: space, la, love, best, graduate
View 8: die, life, black, true, amistad

Movie Name	**Type**
Babysitters club the 1995,	Children
Strip tease 1996,	Comedy | Crime
All Strippers must die,	Horro | Crime
Hangmen also die,	Drama | War

MT-CRL can alleviate spurious correlation

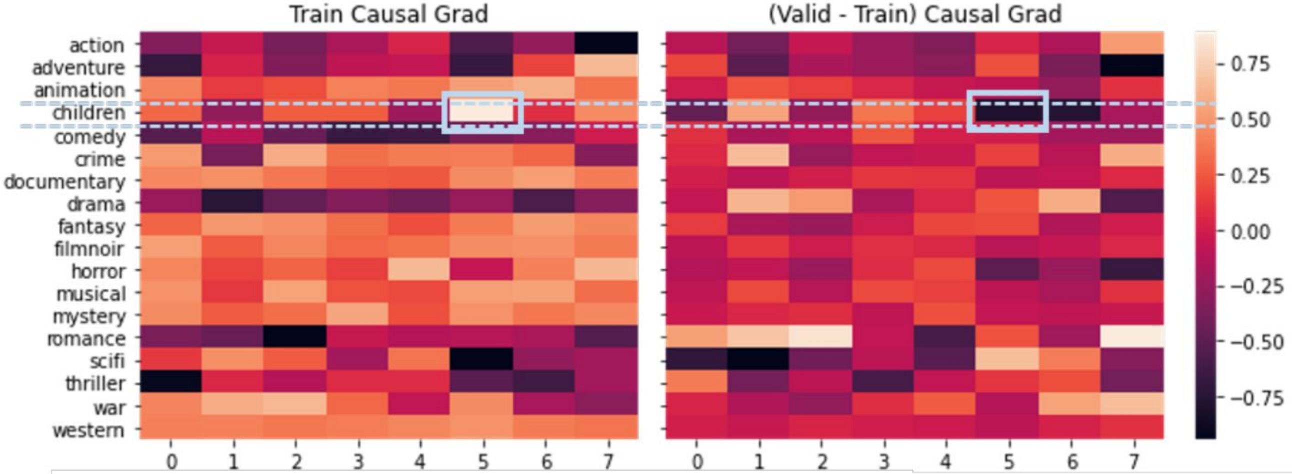


View 1: Shine, go, shawshank, psycho, dumber
View 2: Rocky, october, casino, muppet, payback
View 3: forrest, gump, carrie, now, saving
View 4: i, house, monty, at, life, dark
View 5: good, club, young, stripes, die
View 6: 1978, out, witness, shining, chocolate
View 7: space, la, love, best, graduate
View 8: die, life, black, true, amistad



(Drama) View 0: amadeus, amistad, farewell, thunderball
(Filmnoir) View 1: spartacus, bad, miracle, croupier
(Mystery) View 2: Werewolf, serpico, wrath, hunt
(Romance) View 3: Wives, Sister, Guys, Titanic
(Children) View 4: Pink, Parenthood, Alice, Jungle
(Animation) View 5: Titans, apollo, dancing, willy
(Musical) View 6: singers, chuck, arlington, lovers
(Adventure) View 7: cube, walking, benjamin, felicia

Without MT-CRL (baseline):



Top `**children**' words **w/o MT-CRL**:

good, club, young, strip, die

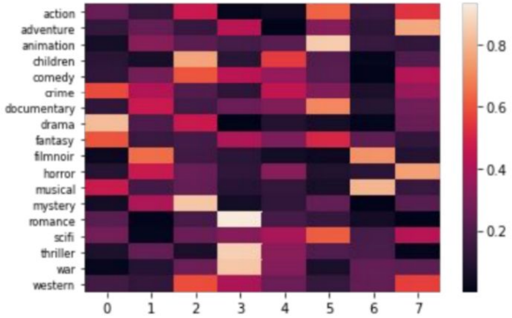
'baby sitters **club** the 1995', **Children**

'**Strip** tease 1996', Comedy|**Crime**

'All **Strippers** must **die**', Horror|**Crime**

'hangmen also **die**, 'Drama|**War**'

With MT-CRL:



(Drama) View 0: amadeus, amistad, farewell, thunderball

(Filmnoir) View 1: spartacus, bad, miracle, croupier

(Mystery) View 2: Werewolf, serpico, wrath, hunt

(Romance) View 3: Wives, Sister, Guys, Titanic

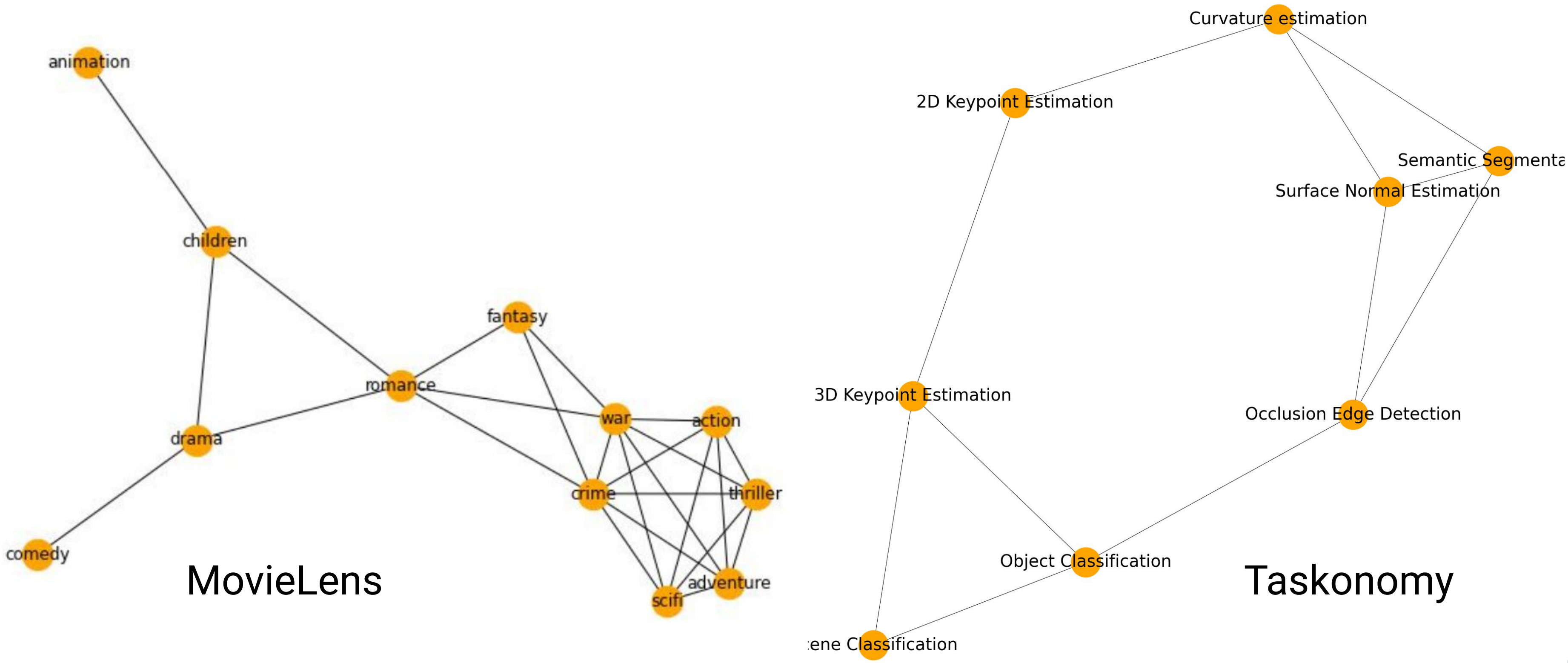
(Children) View 4: Pink, Parenthood, Alice, Jungle

(Animation) View 5: Titans, apollo, dancing, willy

(Musical) View 6: singers, chuck, arlington, lovers

(Adventure) View 7: cube, walking, benjamin, felicia

MT-CRL can learn cross-task similarity



Thanks for Listening~