南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

In Pursuit of Causal Label Correlations for Multi-label Image Recognition

Zhao-Min Chen[1], **Xin Jin**[2], **Yisu Ge**[1,*], **Sixian Chan**[3]
[1]Key Laboratory of Intelligent Informatics for Safety & Emergency of Zhejiang Province, Wenzhou University
[2]Samsung Electronic (China) R&D Centre, Samsung Electronic
[3]The College of Computer Science and Technology, Zhejiang University of Technology
chenzhaomin123@gmail.com, ysg@wzu.edu.cn

NeurIPS 2024

Introduction

- Multi-label image recognition is a fundamental task in computer vision, aiming to predict all objects present in an image.

- However, this task is challenging as the combinations of labels can be tremendous. Modelling label correlations to reduce the search space

- An common assumption is: the training and test sets follow independent and identically distributions (i.i.d.), and the label correlations are consistent. graph structures or attention mechanisms have been successfully employed.

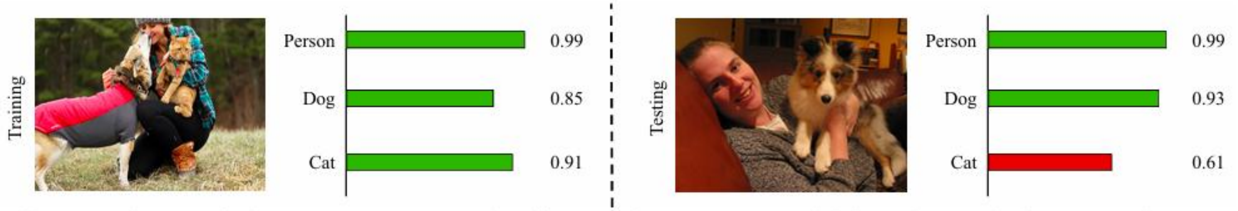

Figure 1: Illustration of the concept and effect of contextual bias in training. It is common that "Person", "Dog", and "Cat" co-occur in training images (we only show one image), while the test image may only contain "Person" and "Dog". Excessive reliance on the label co-occurrence in the training set may lead the recognition model to predict the "Cat" solely based on the presence of "Person" and "Dog".

Introduction

a few researchers attempted to alleviate the effects of contextual bias by

- decorrelating the feature representations of a category from its co-occurring context
- removing the contextual bias in features with causal mechanisms.

We pursue causal correlations (e.g., from “Person” to “Clothes”) to mine contextual cues for recognition, while suppressing spurious correlations (e.g., from “Person” to “Cat”) which are associated by confounders (e.g., the overall scene)

Motivation of Causal Correlations

- a calculable formal definition： if P(Y |do(X)) > P(Y), then a causal correlation exists from X to Y in a probability-raising sense.

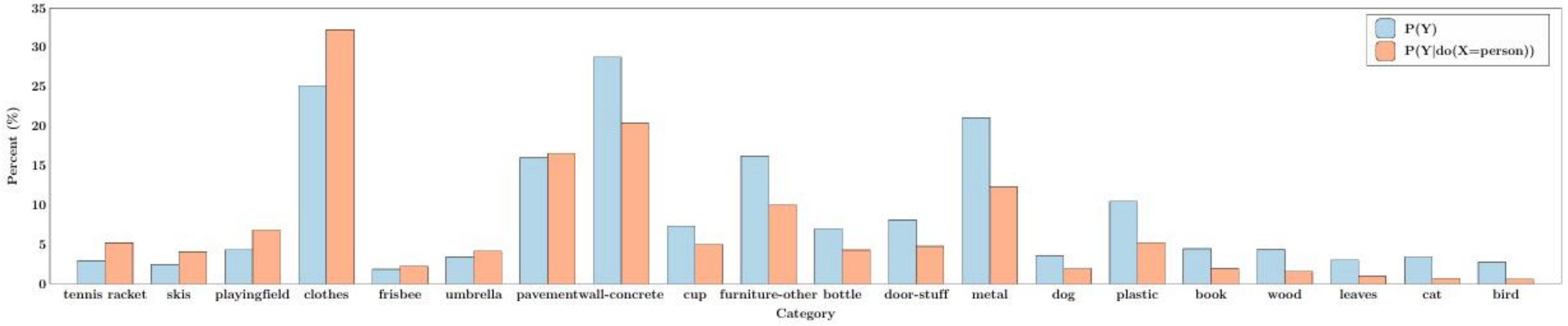

Figure 2: Illustration of causal label correlations and spurious correlations revealed by causal intervention, in a probability-raising sense that if $P(Y|do(X)) > P(Y)$, then a causal correlation exists from X ("Person") to Y (categories in this figure).

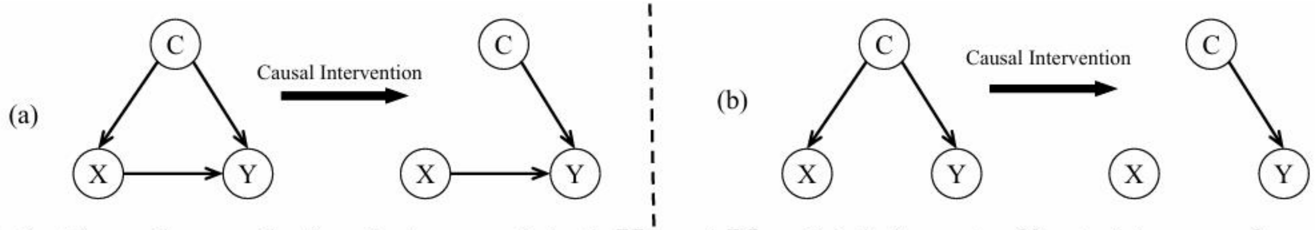

Figure 3: (a): Causal correlation between label X and Y, which is not affected by confounder set C. (b): Spurious correlation, where the co-occurence of X and Y is caused by confounder set C.

Motivation of Causal Correlations

- backdoor adjustment: $$P(Y|do(X)) = \sum_{c} P(Y|X, C=c)P(C=c)$$

As “physical” intervention that puts Y at any context is almost impossible, backdoor adjustment is typically applied for “virtual” intervention.

Here the key idea is to cut off the link from confounder C to cause X, and stratify C into pieces C ={c},making C no longer correlated with X.

Approach

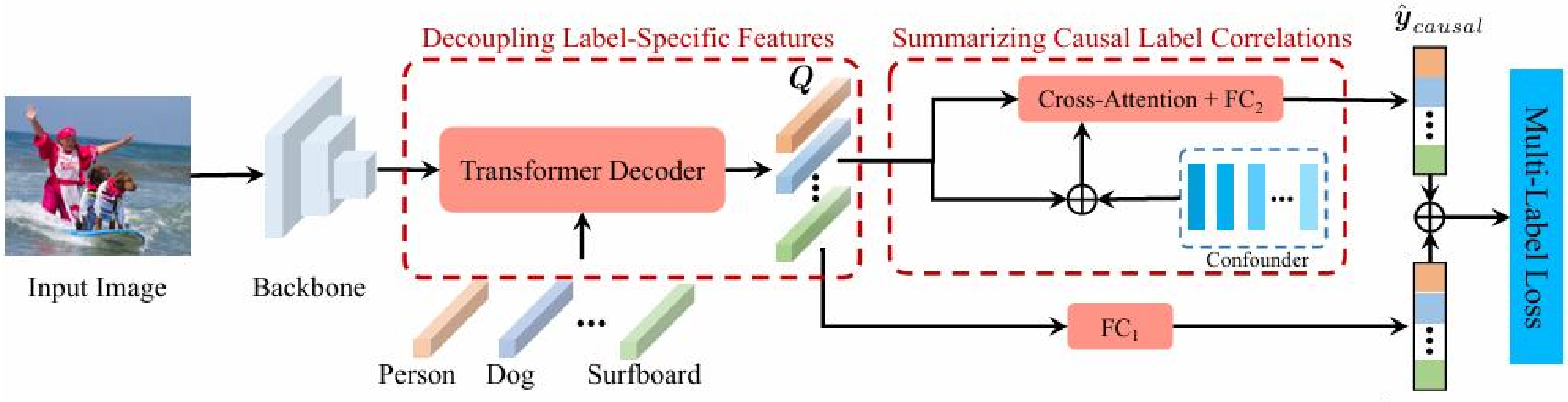

Figure 4: The overall framework of our proposed method.

(a)the branch of decoupled label-specific features
(b)the branch of causal label correlations.

- final prediction confidence score: $\hat{\boldsymbol{y}} = 1/2 \cdot \hat{\boldsymbol{y}}_{causal} + 1/2 \cdot \hat{\boldsymbol{y}}_{decouple} \in \mathbb{R}^N$
- final Loss: $\mathcal{L} = \sum \boldsymbol{y}_{gt} \log(\hat{\boldsymbol{y}}) + (1 - \boldsymbol{y}_{gt}) \log(1 - \hat{\boldsymbol{y}})$

where $\boldsymbol{y}_{gt} = \{0, 1\}^N$ is ground truth label vector of the input image.

Approach

Decoupling Label-Specific Features

$$\boldsymbol{F} = f_{cnn}(\boldsymbol{I}) \qquad \boldsymbol{X} = f_{decoder}(\boldsymbol{Q}, \boldsymbol{F}) \qquad \hat{\boldsymbol{y}}_{decouple} = \sigma(f_{fc1}(\boldsymbol{Q}))$$

Causal intervention based on label-specific features

$$P(Y_j|X_i, C=c) = \sigma(f_{y_j}(x_i, c))$$

$$\begin{aligned} P(Y_j|do(X_i)) &= \mathbb{E}_c[\sigma(f_{y_j}(x_i, c))] \\ &\approx \sigma(\mathbb{E}_c[f_{y_j}(x_i, c)]) \\ &= \sigma(\sum_c f_{y_j}(x_i, c) \cdot P(c)) \end{aligned}$$

Effective modeling for all fyj(xi,c) by cross-attention

let $\boldsymbol{X} = [x_1, ..., x_N] \in \mathbb{R}^{N \times D}$ denote all label-specific features

$$\begin{aligned} \boldsymbol{Z}_c &= \boldsymbol{X} + c\,, \\ \hat{y}^j_{causal} &= f_{merge}([P(Y_j|do(X_1), ..., P(Y_j|do(X_N)]) \\ &= f_{merge}([\sigma(\sum_c f_{y_j}(x_1, c) \cdot P(c)), ..., \sigma(\sum_c f_{y_j}(x_N, c) \cdot P(c))]) \\ &\approx \sigma(\sum_c f_{y_j}(\boldsymbol{X}, c) \cdot P(c)) \\ &= \sigma(\sum_c f_{fc2}(f_{cross\_atten}(y_j, \boldsymbol{Z}_c, \boldsymbol{Z}_c)) \cdot P(c))\,, \end{aligned}$$

Approach

Modeling the confounders

clustering spatial features with K-means algorithm, we obtain a compact set of prototypes to represent potential confounders like objects, scenes and textures.

Experimental results

CommonSetting

- "Exclusive" denotes virtual co-occurrence where labels appearing simultaneously in the training set do not co-occur in the test set.
- "Co-occur" represents the objects co-occurring in both the training and test sets.
- "All" is the average performance of all categories.

Table 1: Performance comparison in the common setting on the COCO-Stuff and DeepFashion datasets.

Method	COCO-Stuff (mAP)			Deepfashion (top-3 recall)		
	Exclusive	Co-occur	All	Exclusive	Co-occur	All
Q2L [19]	23.5	67.1	57.2	12.8	26.3	26.1
ADD-GCN [12]	20.6	64.8	55.2	8.2	22.6	23.5
ML-GCN [4]	18.6	67.1	55.1	10.3	23.7	24.0
SSGRL [28]	18.1	66.6	54.9	7.9	22.8	23.1
C-Tran [15]	22.4	65.1	55.4	11.4	24.6	24.8
CCD [17]	23.8	65.3	55.9	11.5	24.2	24.6
TDRG [38]	20.0	64.8	56.2	8.1	22.9	23.6
IDA [18]	25.2	64.9	57.0	11.3	25.1	25.4
CAM-Based [14]	26.4	64.9	–	–	–	–
feature-split [14]	28.8	66.0	–	9.2	20.1	–
Baseline (R50)	21.9	65.5	55.0	11.5	24.1	24.1
Ours	**29.7**	**69.6**	**60.6**	**14.6**	**27.4**	**28.8**

Experimental results

Cross-dataset Setting

Table 2: mAP Performance comparison in the cross-dataset setting on the MS-COCO and NUS-WIDE datasets.

Method	MS-COCO $\rightarrow$ NUS-WIDE	NUS-WIDE $\rightarrow$ MS-COCO
ADD-GCN [12]	81.8	77.2
ML-GCN [4]	81.4	77.2
SSGRL [28]	80.2	76.1
C-Tran [15]	80.9	76.9
CCD [17]	81.9	78.3
Q2L [19]	82.1	78.6
IDA [18]	82.3	78.9
CAM-Based [14]	81.0	77.8
feature-split [14]	81.9	78.3
Baseline (R101)	81.1	77.1
Ours	**83.2**	**80.2**

Ablation Studies

Table 3: The impacts of different modules.

Decouple	Causal	Exclusive	Co-occur	All
		21.9	65.5	55.0
√		22.1	67.0	56.8
√	√	**29.7**	**69.6**	**60.6**

Table 4: The impacts of clustering center number.

Number	Exclusive	Co-occur	All
20	26.9	68.9	59.6
40	26.8	69.3	60.1
60	29.3	**69.8**	60.2
80	**29.7**	69.6	**60.6**
100	29.5	69.4	60.5

Table 5: The impact of backbones for clustering.

Confounder Backbone	Exclusive	Co-occur	All
ResNet-50	**29.7**	69.6	**60.6**
ResNet-101	29.6	**69.9**	60.5
BEIT3-Large	29.4	69.7	60.5

Table 6: The impact of different modeling approaches for confounders.

Method	Exclusive	Co-occur	All
Random	22.1	66.3	56.1
Early	27.8	69.1	60.1
Label	28.0	69.3	60.3
K-means	**29.7**	**69.6**	**60.6**

Table 7: Different implementations of Eq. $\boxed{7}$.

Method	Exclusive	Co-occur	All
Linear	27.4	69.1	60.0
Ours	**29.7**	**69.6**	**60.6**

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

THANKS