Debiasing Event Understanding for Visual Commonsense Tasks

Minji Seo[1]*, **YeonJoon Jung**[2]*, **Seungtaek Choi**[2], **Seung-won Hwang**[1]†, **Bei Liu**[3]

[1]Seoul National University, [2]Yonsei University, [3]Microsoft Research Asia
{minjiseo, seungwonh}@snu.ac.kr, {theaitetus, hist0613}@yonsei.ac.kr,
Bei.Liu@microsoft.com

Abstract

We study event understanding as a critical step towards visual commonsense tasks. Meanwhile, we argue that current object-based event understanding is purely likelihood-based, leading to incorrect event prediction, due to biased correlation between events and objects. We propose to mitigate such biases with *do*-calculus, proposed in causality research, but overcoming its limited robustness, by an optimized aggregation with association-based prediction. We show the effectiveness of our approach, intrinsically by comparing our generated events with ground-truth event annotation, and extrinsically by downstream commonsense tasks.

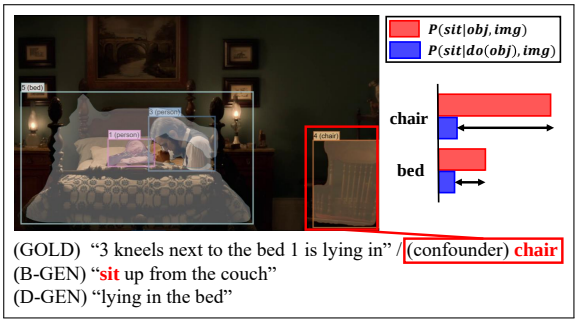

Figure 1: An actual example of event generation. By removing the chair (`confounder`), the biased generation (`B-GEN`) can be corrected to the true event (`D-GEN`).

1 Introduction

Recently, commonsense reasoning tasks on visio-linguistic input have been actively studied in both vision and language communities, with the goal of commonsense reasoning. For example, in Visual COMET (Park et al., 2020), given an image X and event e as input, we are tasked to predict intent (or, events before/after), known as **intent** prediction. Similarly, Visual Commonsense Reasoning (VCR; Zellers et al. 2019), given image and question Q, requires to provide an answer A or provide a rationale for A, known as **justification** task.

For such tasks, understanding event e plays a crucial role, either required as input or expected as output–for example, 41% of rationales from VCR are related to e. However, assuming the availability of e for new images at test time is considered impractical, for which Park et al. (2020) consider two baselines: 1) setting input e as `NULL`, and 2) generating e by training generator `GEN`. However, according to Park et al. (2020), empirical results reported `GEN` performs even worse than `NULL`, or, shows a negative transfer, which is counter-intuitive.

Our goal is robustifying event understanding, or, event generator `GEN` with a higher **transferability** to the downstream tasks of visual commonsense reasoning. Our hypothesis is that object-event bias in the dataset hinders transferability: In Figure 1, while the image and human-annotated e, denoted as `GOLD`, are not related to an event "sitting", existing `GEN`, denoted as `B-GEN` for Biased GENeration, generates an event "sitting", even when no one is sitting on the chair. We argue that the generation is biased by the frequent **association** of chair-sitting in the dataset.

In contrast, we propose `D-GEN` for Debiased GENeration. Contrary to association-based estimation, which fails to distinguish spurious correlation $P(\text{sitting}|\text{chair})$, *do*-calculus collectively considers other observations, such as "sitting" on other objects (*e.g.*, bed or table), to lower the likelihood of such spurious correlations as humans do. However, while *do*-calculus has been successful at debiasing other types of biases, e.g., between word and visual word (Zhang et al., 2020), we find a limitation for object-event debiasing as in Figure 2: Though *do*-calculus can find a causal relation, *i.e.*, diningtable-sit, with high $P(\text{sit}|do(\text{diningtable}))$, the same logic does not apply to rarely observed ob-

*Equal contribution
†Corresponding author

Findings of the Association for Computational Linguistics: ACL 2022, pages 782 - 787

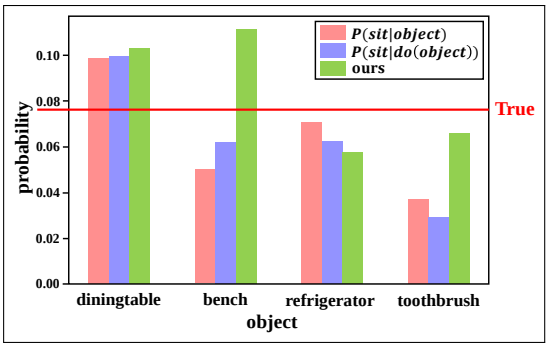

Figure 2: Comparative example of $P(Y|X)$, $P(Y|do(X))$, and ours in VisualCOMET dataset. These are actual estimation results from our model.

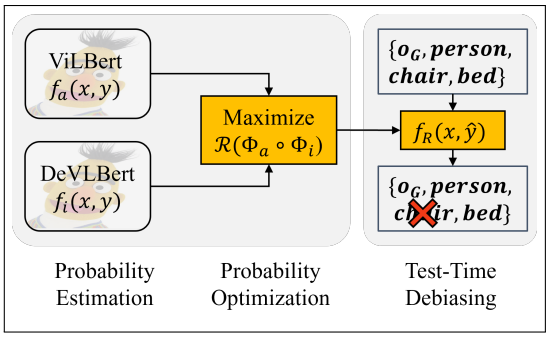

Figure 3: Framework overview.

ject 'bench'. In other words, $P(\text{sit}|do(\text{bench}))$ and $P(\text{sit}|do(\text{refrigerator}))$ are similar, where we cannot easily distinguish the true causal relation, *i.e.*, bench-sit, from non-causal one, *i.e.*, refrigerator-sit.

To motivate the needs to robustify do, Figure 2 enumerates four representative classes, where $P(Y|X)$ (and $P(Y|do(X))$) is high/low respectively: the first two objects (diningtable and bench) have a causal relation with event 'sit'. While do (pink bar in the figure) can only identify the first object, we propose an ensemble (green bar) of both probabilities, which can distinguish the first two from the rest. Specifically, we search over a space Ω of possible aggregations of the two, from which we identify a robust causality scoring f_R. Contrary to do-calculus, our estimation better distinguishes frequent-causal correlation, *i.e.*, diningtable-sit, as well as rare-causal correlation, such as bench-sit, from the remaining two.

To apply the robust estimator trained as above for a test image without event, we deliberately instill the biases by biased event generation, which we are trained to remove, as inspired by Qian et al. (2021). We then mitigate biases in the testing time, by generating counterfactual image $\hat{X}_R$, eliminating the confounder identified from f_R.

Experimental results on VisualCOMET and VCR show that our method significantly improves the robustness of visual commonsense models and our codes can be found from supplementary material.

2 Methodology

Figure 3 overviews our framework in three steps: (1) probability estimation, (2) probability optimization, and (3) test-time debiasing.

2.1 Probability Estimation

As discussed above, we propose a new estimation of causality by combining both $P(Y|do(X))$ and $P(Y|X)$. For this purpose, we first leverage DeVLBert (Zhang et al., 2020) to estimate $P(Y|do(X))$, whose pre-training task, of leveraging 3.04 million <image, caption> to debias spurious correlations between image and words, is relevant to our objective of debiasing object-event. We thus aim to transfer the pre-trained DeVLBert to fit in our problem, by a fine-tuning task that can adapt to words that are important for event debiasing, e.g., verb. Specifically, we train DeVLBert with a classification task of matching image X with the right event description Y, generating a high score for a matching $X-Y$ pair.

In our classification task, DeVLBert takes the image X and the textual event description Y separately. Specifically, for textual input, it follows Bert convention (Devlin et al., 2018), where each token is represented as a sum of its corresponding token/position/segment embeddings, yielding a sequence of embeddings, *i.e.*, $y = \{w_{[CLS]}, w_1, ..., w_{|Y|}\}$. The special token [CLS] is used to capture the global information in the text. For visual input, by viewing the sub-regions of interest as visual words, the image is represented as a sequence of visual words, where each visual word (object) is detected by Faster-RCNN (Anderson et al., 2018) and the features of visual words $\{o_i\}_{i=1,...,k}$ are extracted by ResNet101 (He et al., 2016). Similarly with textual input, a global representation for the whole image $o_{[G]} = \frac{1}{k}\sum_{i=1}^{k} o_i$ is added at the beginning of the sequence, resulting $x = \{o_{[G]}, o_1, \ldots, o_k\}$. By finally feeding x and y into co-attentional transformer layers (Lu et al., 2019), DeVLBert is trained to predict the gold event description y among candidate event descriptions that are randomly sampled as negative.

Once this training is done, we can reconstruct $P(Y|do(X))$ by computing the fine-tuned DeVLBert scores $f_i(x, y)$ for event candidates, and normalizing the scores into a probability distribution.

Then, we adopt ViLBert (Lu et al., 2019) to estimate $P(Y|X)$, which we denote as f_a, following the same protocol described above. We stress that our proposed solution, building upon ViLBert/DeVLBert, is orthogonal to other task models using Faster-RCNN features (Anderson et al., 2018) such as UNITER and VILLA (Chen et al., 2020; Gan et al., 2020).

2.2 Probability Optimization

As we claimed above, it is important to distinguish rare-but-causal correlations from frequent-but-spurious correlations, for which we propose a more robust estimation f_R by aggregating $P(Y|X)$ and $P(Y|do(X))$. In this work, we formalize it as an optimization problem, searching over a space Ω of possible aggregations of $P(Y|X)$ and $P(Y|do(X))$, for the goal of maximizing an objective function $\mathcal{R}$.

We thus aim to enumerate the search space Ω to maximize the objective function $\mathcal{R}$. However, as exhaustively enumerating the search space Ω is infeasible, we consider the following two desired properties for f_R that can reduce the search space: 1) Positive correlation with f_i to preserve the strength of $P(Y|do(X))$ for identifying frequent-and-causal correlations, and 2) Negative correlation with f_a to prevent the over-estimation problem of $P(Y|X)$ for frequent-but-spurious correlations, where the optimization process for the robust estimation of causality f_R can be written as follows:

$$\underset{\Phi_a,\Phi_i,\odot}{\text{maximize}} \quad \mathcal{R}(\Phi_a(f_a) \odot \Phi_i(f_i)), \tag{1}$$

$$\text{s.t.} \quad \Phi_a(f_a) = \lambda_a \ominus f_a, \tag{2}$$

$$\Phi_i(f_i) = \lambda_i \oplus f_i, \tag{3}$$

$$\lambda_a, \lambda_i \in \{-1, 0, 1\}, \oplus \in \{+, *\}, \tag{4}$$

$$\ominus \in \{-, /\}, \text{and } \odot \in \{+, *\}, \tag{5}$$

where the objective function $\mathcal{R}$, to directly examine the effectiveness of f_R regarding its capability of identifying causal correlations, is defined as the number of correctly identified spurious correlations, which can be validated with human-selected confounders (Section 3).

2.3 Test Time Debiasing

Now that we trained a robust confounder identifier f_R, the next step is applying f_R on the test images to explicitly eliminate confounders. However, as the event y is missing for test image, we predict an event $\hat{y}$ by employing "poisonous" model with the same bias, namely f_a, such that we use $\hat{y} = \arg\max_{y \in \mathcal{E}} f_a(x, y)$[1]. As f_R is trained to distill biases from association-based prediction, with the predicted event $\hat{y}$, we can identify the confounder object o_c with the lowest causality, likely spurious: $o_c = \arg\min_{o_i} f_R(x_{o_i}, \hat{y})$, where x_{o_i} denotes a constrained input for measuring the causality of a single object o_i, *i.e.*, $x_{o_i} = \{o_{[G]}, o_i\}$. Note that this selection can be iterated with a score threshold (tuned from validation set) for multi-confounder cases. Without loss of generality, we consider o_c as a single object in this section.

With the identified confounder o_c, our final goal is to obtain a debiased image $\hat{X}_R$ for the purpose of robust event understanding, or event generator `GEN` with high transferability. To this end, we propose to remove all the visual features of objects that are of the same class with the identified confounder: $\hat{X}_R = X \setminus \{o_i | o_i \approx o_c\}$. By feeding the debiased input $\hat{X}_R$ for downstream tasks, our method will mitigate the spurious correlations in task models.

3 Experimental Settings

Dataset	Train	Valid	Test
VisualCOMET	47.5K	5.9K	5.9K
VCR	80.4K	9.9K	9.5K

Table 1: Statistics for datasets.

To evaluate the effectiveness of our framework, we conduct experiments on VisualCOMET and VCR datasets, by training model on the former, and study out-of-domain generalization for the latter. We report results for event, intent, and rationale prediction tasks from each dataset, respectively. The statistics of the datasets are presented in Table 1.

To adopt ViLBert/DeVLBert for f_a and f_i, we trained ViLBert/DeVLBert with our classification task for 20 epochs with a batch size of 64, and the initial learning rate is set as 2e-5. We adopt standard models for each tasks. For event understanding and intent inference tasks, We adopt GPT-2 based

[1]We constraint $\mathcal{E}$ the 111,796 training set events provided in (Park et al., 2020).

single-stream Transformer architecture, introduced in Park et al. (2020). Following the convention in Park et al. (2020), the parameters are optimized by Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5e-5 and batch size of 64. For the VCR-justification task, we adopt ViLBert (Lu et al., 2019), which takes the concatenated question and correct answer as query to predict rationales. Following the settings of ViLBert, we fine-tuned ViLBert by Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2e-5 and batch size of 64. For both tasks we used the maximum number of detected objects k is set as 15.

To build human-selected confounders for the validation purpose (Section 2.2), we first present the objects to human workers in the order of frequency, then ask them to filter out the objects with causality, so that confounders, or, spurious object-event pairs remain. We stress that the annotation process is efficiently guided by machine selections, such that humans are not exposed to all pairs, but only the surviving pairs from the first-phase machine selection, to filter out objects with causality, such that only the spurious object-event pairs remain after the second phase. We also found causality can be reliably annotated among multiple annotators, from a substantial inter-annotator agreement: Human-selected confounders agreement was measured to be 0.689 in Cohen's Kappa coefficient.

4 Experimental Results

We now proceed to empirically validate the effectiveness of our approach, in three dimensions: 1) capturing rare-but-causal correlations 2) transferability and 3) out-of-domain generalization.

Capturing Rare-but-causal: To investigate the effectiveness of our approach in capturing rare-but-causal correlations, we perform the event understanding task on the VisualCOMET dataset. For the experiment, we split the validation samples into a frequent set and a rare set, where the latter requires identifying rarely observed causal correlations. Among the images with the rarely observed object-verb pairs, *i.e.*, occurrences < 20, we collect the images with an object-verb, where the verb exists in the ground-truth event description E_{gold}, as rare-but-causal set. A desirable model should generalize well to the challenging set, or, the gap between the original and challenging sets should be small.

We report our results on event understanding in Table 2, reporting METEOR (Denkowski and Lavie, 2014) between the generated events ($E_{gen}(\hat{X}_I)$, $E_{gen}(\hat{X}_R)$) with the gold event E_{gold} on frequent set, rare set, and total set respectively. As a baseline, we compare ours with the intervention-based baseline $\hat{X}_I$. In frequent set, $\hat{X}_I$ and $\hat{X}_R$ showed similar results as both are capable of capturing causal correlations when they appear frequently. On the contrary, in the case of rare-but-causal correlations, the gap between ours $\hat{X}_R$ and $\hat{X}_I$ increases, showing the strengths of ours at distinguishing rare-but-causal correlations from frequent-but-spurious correlations. As a validation of such strengths, we compare the error rate on detecting rare-but-causal correlation of f_R and f_i. We confirmed that f_R significantly reduces the error rate from 19% of f_i to 15%, verifying that the event understanding benefits from distinguishing rare-but-causal correlation out of frequent-and-spurious correlation.

Method	Frequent	Rare	Total
$E_{gen}(\hat{X}_I)$	**14.16**	12.29	13.50
$E_{gen}(\hat{X}_R)$	14.15	**12.97**	**13.74**

Table 2: Results on event understanding task.

Image	Event	BLEU-2	METEOR	CIDEr
X	NULL	2.14	4.51	3.76
X	$E_{gen}(X)$	2.26	4.52	4.40
$\hat{X}_R$	$E_{gen}(\hat{X}_R)$	**2.90**	**4.67**	**5.93**
$\hat{X}_R$	E_{gold}	2.95	5.63	8.66

Table 3: Results on intent inference task. We compare each method for randomly sampled 100 validation examples.

Transferability: We evaluate how our approach contributes to transferability, which we define as the gap between event description as NULL and GEN, on intent inference task. We report our results on the intent inference task in Table 3, reporting BLEU-2 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), and CIDEr (Vedantam et al., 2015) between the inferred intent with the gold intents. Specifically, we extend the comparisons with input debiasing for the generated events, *i.e.*, $E_{gen}(X) \rightarrow E_{gen}(\hat{X}_R)$. In the table, we observe that, without input debiasing, generated event $E_{gen}(X)$ is not beneficial, achieving only a minor improvement in BLEU-2 (2.14 $\rightarrow$ 2.26), due to its low transferability as Park et al. (2020) observed a decrease. On the other hand,

we observe that when $E_{gen}(\hat{X}_R)$ is equipped with the debiased input $\hat{X}_R$, the model performs best (performing closely to the oracle case with human annotation GOLD), showing the effectiveness of debiasing towards transferability.

Out-of-domain Generalization: A reliable neural network should be able to generalize across the distribution shift, *i.e.*, test distribution is different from the training distribution. To make the point that our approach can make the model generalize over distribution shifts, we conduct experiments on the VCR-justification task, where the linguistic parts may not overlap with VisualCOMET.

	X	$\hat{X}_I$	$\hat{X}_R$
Top-1 Acc (%)	62.94	63.90	**64.22**

Table 4: Results on justification task in VCR dataset.

The results of the VCR-justification task are presented in Table 4. In accordance with the result of former experiments, the proposed debiased input $\hat{X}_R$ significantly improves the result without further training, achieving 1.28 point accuracy gain from $X \rightarrow \hat{X}_R$, while baseline input debiasing $\hat{X}_I$ achieving only 0.96 point gain. It demonstrates that our proposed input debiasing can generalize on out-of-domain tasks by better distinguishing confounder objects of spurious correlation.

5 Related Work

Inspired by the success of large-scale pre-trained language models (Devlin et al., 2018; Radford et al.), large-scale pre-training on transformers have shown that it can also benefit visio-linguistic tasks, showing better transferability on various downstream tasks (Lu et al., 2019; Li et al., 2020). However, it is reportedly bad when the transformer is trained on out-of-domain datasets, that are not aligned with its pre-training corpus (Chen et al., 2020; Zhang et al., 2020), as they are purely likelihood-based, leading to spurious correlations and hurting the generalization ability. To this end, recent approaches adopt a traditional *do*-calculus (Pearl et al., 2016), encouraging causal intervention-based estimation to remove spurious correlations, that can be exploited in vision-only dataset (Wang et al., 2020) or visio-linguistic dataset (Zhang et al., 2020; Yang et al., 2021). Different from these works, we focus on identifying spurious correlation comparing association-based and intervention-based knowledge.

6 Conclusion

We studied the problem of robustifying event understanding, to overcome dataset bias, by combining observational and interventional estimations. Our experiments suggest that this extension improves event understanding, and eventually visual commonsense tasks.

Acknowledgements

This research was supported by Microsoft Research Asia, SNU-NAVER Hyperscale AI Center, and IITP grants funded by the Korea government (MSIT) [2021-0-02068 SNU AIHub, IITP-2022-2020-0-01789].

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 6616–6628. Curran Associates, Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer.

J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal attention for vision-language tasks.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382.