DeVLBert: Out-of-distribution Visio-Linguistic Pretraining with Causality

Shengyu Zhang[1*], Tan Jiang[1*], Tan Wang[2], Kun Kuang[1†], Zhou Zhao[1], Jianke Zhu[1], Jin Yu[3], Hongxia Yang[3†], Fei Wu[1†]

[1] Zhejiang University, [2] University of Electronic Science and Technology of China, [3] Alibaba Group

{sy_zhang, jiangtan, zhaozhou, kunkuang, jkzhu, wufei}@zju.edu.cn

{kola.yu, yang.yhx}@alibaba-inc.com, wangt97@hotmail.com

Abstract

In this paper, we propose to investigate out-of-domain visio-linguistic pretraining, where the pretraining data distribution differs from that of downstream data on which the pretrained model will be fine-tuned. Existing methods for this problem are purely likelihood-based, leading to the spurious correlations and hurt the generalization ability when transferred to out-of-domain downstream tasks. By spurious correlation, we mean that the conditional probability of one token (object or word) given another one can be high (due to the dataset biases) without robust (causal) relationships between them. To mitigate such dataset biases, we propose a Deconfounded Visio-Linguistic Bert framework, abbreviated as DeVLBert [1], to perform intervention-based learning. We borrow the idea of the backdoor adjustment from the research field of causality and propose several neural-network based architectures for Bert-style out-of-domain pretraining. The quantitative results on three downstream tasks, Image Retrieval (IR), Zero-shot IR, and Visual Question Answering, show the effectiveness of DeVLBert by boosting generalization ability [2].

1. Introduction

Since early attempts that pretrain a backbone model on large-scale dataset and then transfer the knowledge to numerous vision and language tasks, pretraining has become a hallmark of the success of deep learning. Despite the significant progress that recent methods have made over the initiative work ViLBert [6], part of their success can be traced back to the introduction of *in-domain* pretraining datasets besides the Conceptual Caption [8] dataset. By *in-domain*, we refer to those datasets used in both pretraining and downstream tasks. However, out-of-domain pretraining, *i.e.*, pretraining models on *out-of-domain* datasets and transferring the learned knowledge into downstream tasks with **unkown** data distributions, can be an essential research topic. In this paper, we focus on out-of-domain pretraining and learning generic representations as the ViLBert does.

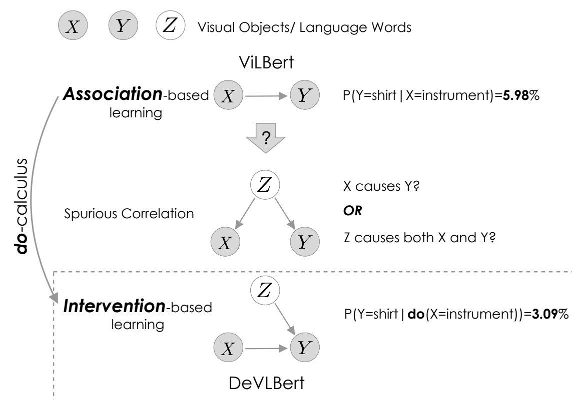

Figure 1. An illustration of the transition from traditional association-based learning to causal intervention-based learning.

A fundamental requirement for out-of-domain transfer learning is to mitigate the biases from the pretraining data [9], which may be useful for the in-domain testing but harmful for out-of-domain testing [2] due to the *spurious correlation*. Most previous works just blame this for the biased data collection without further justification. However, this is not reasonable since we human ourselves are just living in a biased nature. In our methodology, we draw inspiration from the causal inference [2] and borrow the idea of the backdoor adjustment (also known as covariate adjustment or statistical adjustment) [7] to mitigate these biases. The essence of deconfounding with the do-operation can be found in Figure 1. In this way, the pure *association*-based pretraining becomes to the causal *intervention*-based pretraining. We are particularly targeting at the Bert-style pretraining models and the context-based proxy tasks for supervision, such as masked language/object modeling (MLM/MOM). Context-based proxy tasks solely care about association, *i.e.*, what co-occur with the anchor to-

[1]Please refer to the full version of this paper [11] for better clarity.
[2]https://github.com/shengyuzhang/DeVLBert
*These authors contributed equally to this work.
†Corresponding Authors.

ken without considering whether there are spurious correlations (*e.g.*, `shirt` cannot cause `instrument`, and vice versa) or not. We propose several intervention-based BERT architectures to help learn deconfouned visio-linguistic representations. We name this kind of architectures as **DeVL-Bert**, which refers to ***Deconfounded Visio-Linguisitic Bert***. DeVLBert is designed as model-agnostic and can be easily encapsulated into any other Bert-style models.

We conduct in-depth experiments to discuss the performance of the proposed DeVLBert architectures. Out-of-domain pretrainin with three downstream vision-language tasks demonstrate that DeVLBert can boost the generalization ability by mitigating dataset biases.

2. Deconfounded Vsio-Linguistic Bert

2.1. Bert in the causal view

As illustrated in [11], the Transformer layer connects each output token representation with all input token representations. We denote the representation of one output token as Y and the representations of all other tokens as X. Bert models the function of $P(Y|X)$. In the causal view, there can be some confounder Z affecting both X and Y. Formally, by the Bayes Rule, the conventional likelihood can be re-written as:

$$P(Y|X) = \sum_z P(Y|X,z)\underline{P(z|X)}, \tag{1}$$

By using the *do*-calculus [7], we remove any incoming influence to the intervened variable, *i.e.*, X:

$$P(Y|do(X)) = \sum_z P(Y|X,z)\underline{P(z)}. \tag{2}$$

It is infeasible to individually model the distribution of $P(Y|X,z)$ for each z as the number of potential confounders can be large. We borrow the idea of Normalized Weighted Geometric Mean [10] to approximate the expensive sampling. Formally, if the last objective is classification, we can re-write the following terms:

$$P(Y|do(X)) = \mathbb{E}_z\left[\sigma\left(f_c(\mathbf{x},\mathbf{z})\right)\right], \tag{3}$$

where $\mathbf{x}$ and $\mathbf{z}$ denote the feature representations of X and z, and f_c denotes the classification head of intervention. σ denotes the softmax function The essence of NWGM is to move the expectation into the operation of softmax:

$$P(Y|do(X)) \overset{\text{NWGM}}{\approx} \sigma\left(\mathbb{E}_z\left[f_c(\mathbf{x},\mathbf{z})\right]\right). \tag{4}$$

In this paper, we model the term $f_c(\mathbf{x},\mathbf{z})$ by the feed-forward neural network $\mathbf{W}_c[\mathbf{x}, \alpha_y(\mathbf{z}) * \mathbf{z}]$, where $[,]$ denotes the concatenation operation and $\alpha_y(\mathbf{z})$ denotes the importance factor that is parameterized by $\mathbf{y}$. Formally, we have:

$$\alpha_y(\mathbf{z}) = \frac{(\mathbf{W}_y\mathbf{y})^T(\mathbf{W}_z\mathbf{z})}{\sum_{v\neq\varsigma}(\mathbf{W}_y\mathbf{y})^T(\mathbf{W}_z\mathbf{v})}, \tag{5}$$

$$P(Y|do(X)) = \sigma(\mathbf{W}_c[\mathbf{x}, \sum_z P(\mathbf{z}) * \alpha_y(\mathbf{z}) * \mathbf{z}]). \tag{6}$$

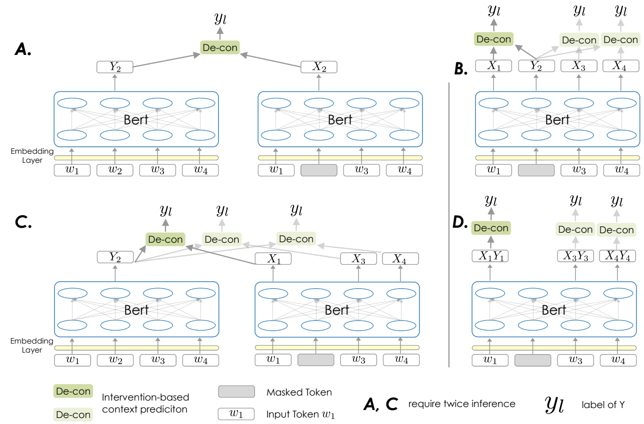

Figure 2. A vivid illustration of four intervention formulations for Bert-style training. Deciding the forms of X and Y in Bert is essential for further intervention-based context prediction. Design A&C require twice inference.

where $\mathbf{y}/\mathbf{v}$ is the feature representation of Y/v. ς denotes the confounder that has the same token class as Y. We propose several implementations for the Bert structure.

Design A. We firstly investigate how to harness masked token modeling with intervention. Still, we take natural language pretraining as an example for illustration. For one masked word w_t, it is intuitively to view the final representation $\mathbf{w}_t$ as $\mathbf{x}_t$ since $\mathbf{w}_t$ contains no explicit information from the word itself (being masked). We choose to run another inference with no masked tokens, and view the final representation of word w_t as $\mathbf{y}$ (shown in Figure 2 A).

Design B. Figure 2 B depicts another design to harness MTM. In this perspective of context modeling, the final representation of the masked token w_t can be viewed as $\mathbf{y}_t$ while the final representations of all unmasked tokens can be viewed as $\{\mathbf{x}_k\}_{k=1,\dots,t-1,t+1,\dots,N_w}$. This design is efficient without an extra inference process. The time complexity is $O(N_u * N_m)$, where N_u and N_m are the numbers of unmasked tokens and masked tokens, respectively.

Design C. As depicted in Figure 2 C, Design C is a variant of Design A and views the final representations of all unmasked tokens as $\{\mathbf{x}_k\}_{k=1,\dots,t-1,t+1,\dots,N_w}$.

Design D. By viewing the final representations of unmasked tokens as integrated representations of X and Y, Design D is non-intrusive and can be the most efficient among the proposed designs. In this design, the modeling of $P(Y|do(X))$ is slightly different:

$$\alpha_r(\mathbf{z}) = \frac{(\mathbf{W}_r\mathbf{r})^T(\mathbf{W}_z\mathbf{z})}{\sum_z(\mathbf{W}_r\mathbf{r})^T(\mathbf{W}_z\mathbf{z})}, \tag{7}$$

$$P(Y|do(X)) = \sigma(\mathbf{W}_c \sum_z P(\mathbf{z}) * \alpha_r(\mathbf{z}) * \mathbf{z}). \tag{8}$$

where $\mathbf{r}$ denotes the integrated representation of $\mathbf{y}$ and $\mathbf{x}$, and $\alpha_r(\mathbf{z})$ is the importance factor parameterized by $\mathbf{r}$. Since the representation $\mathbf{x}$ is no longer available, we omit

Table 1. Comparison between DeVLBert and other competitors, including ViLBERT which only uses out-of-domain$^\circ$ pretraining datasets, VisualBERT only uses in-domain$^\bullet$ datasets, and InterBert using both$^\odot$.

Methods	Image Retrieval (IR)			Zero-shot IR			VQA	
	R@1	R@5	R@10	R@1	R@5	R@10	test-dev	test-std
SCAN [3]	48.6	77.7	85.2	-	-	-	-	-
BUTD [1]	-	-	-	-	-	-	65.3	65.7
$^\bullet$VisualBERT [4]	-	-	-	-	-	-	70.8	71.0
$^\odot$InterBert [5]	61.9	87.1	92.7	49.2	77.6	86.0	70.3	70.6
$^\circ$ViLBERT [6]	58.2	84.9	91.5	31.9	61.1	72.8	70.6	70.9
$^\circ$DeVLBert	61.6	87.1	92.6	36.0	67.1	78.3	71.1	71.5

the concatenation operation.

2.2. Intra- & Inter-modality Intervention

Vision deconfounding & Vision Confounder Set. Following VC R-CNN [9], we consider the high-level object classes as potential confounders. The representation of each object class is obtained by averaging pooling the set of object features belonging to the class (but in different images). For vision deconfounding, Y and X are only selected from the final representations of the visual regions, and confounders in the vision confounder set are discussed. For Design A and B, the MOM objective is totally replaced by the intervention objective. For Design C and D, the intervention is married with the MOM objective.

Language deconfounding & Language Confounder Set. We extract nouns as potential confounders by Part-of-Speech Tagging and filter those of low-frequencies, resulting in 156 potential confounders. The feature representation of each noun is initialized as the mean-pooled vector of the Bert contextual embeddings of words (the same noun) in different sentences. Deconfounding strategy is similar to the vision part.

Inter-modality Intervention. For inter-modality intervention, Y and X can be tokens from different modalities, and confounders can be selected from both vision and language confounder sets.

3. Experiments

Pretraining DeVLBert. We follow ViLBERT [6] to pretrain DeVLBert on the Conceptual Caption [8] dataset, which is an out-of-domain dataset that has little data overlap with most downstream tasks. **Finetuning on downstream tasks.** Also, we are following the pipelines of three downstream tasks, *i.e.*, Text-to-Image Retrieval (IR), Zero-shot Text-to-Image Retrieval (Zero-shot IR), and Visual Question Answering (VQA) of ViLBERT. For more details, such as dataset split, fine-tuning strategies, and hyper-parameters, please refer to ViLBERT[6].

Table 2. Comparisons between different DeVLBert implementations, and ablation studies on the architecture D.

Method	Image Retrieval (IR)			Zero-shot IR		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	58.2	84.9	91.5	31.9	61.1	72.8
A-V	60.3	86.24	92.06	30.18	59.46	71.88
A-VL	58.3	85.5	91.6	25.4	54.7	67.2
B-V	58.9	85.3	91.1	33.0	62.2	74.0
C-V	-	-	-	27.0	56.2	69
D-V	59.3	85.4	91.8	32.8	63.0	74.1
D-VL	60.3	86.7	92.2	34.9	65.5	77.0
D-VLC	**61.6**	**87.1**	**92.6**	**36.0**	**67.1**	**78.3**

3.1. Quantitative Evaluation

How do different intervention-based architectures perform? To answer this question, we evaluate the performance of different architectures on the downstream tasks, *i.e.*, image retrieval, and zero-shot image retrieval. The results are listed in Table 2. We use A-V to denote the architecture of design A, A-VL to denote the architecture of design A with both vision and language deconfounding. Based on the results, we can see that:

1) Most of the architectures obtain performance gain on at least one of the tasks, which demonstrates the effectiveness of intervention-based learning.

2) The twice inference design achieves inferior results on the zero-shot image retrieval task. Due to the complexities introduced by another inference, it might take more iterations to converge, which can be computationally expensive.

3) Comparing A-VL with A-V, the introduction of language deconfounding leads to a performance drop on IR and zero-shot IR. We attribute this phenomenon to the incomplete training of MTM. For the language side, following ViLBERT, the classification module shares the word embedding matrix with the input embedding layer. For A-VL, we only mask noun words since the language confounder set comprises only noun words. Therefore, the embedding matrix solely sees noun words in the classification, which leads to inferior results due to incomplete learning of other words. Non-intrusive design D mitigates this problem.

4) Without the structure and training complexities introduced by the other inference, B-V and D-V show clear advantages over A-C and C-V.

5) D-V further outperforms the architecture of B-V, and we attribute this consistent improvement to the non-intrusive intervention modeling. More concretely, isolating the masked token modeling makes the shared embedding module in the MTM classification module learn better. Meanwhile, architecture D is the most efficient.

Do both intra-/inter- modality intervention improve the out-of-domain pretraining? Since architecture D-V achieves the best performance, we further extend architecture D-V to architecture D-VL by incorporating language

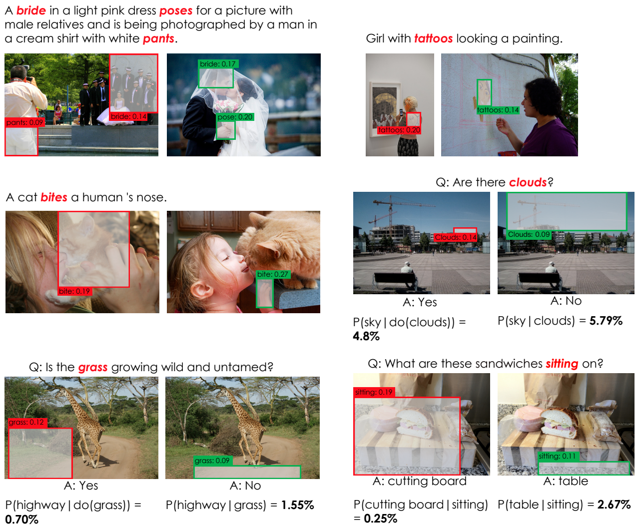

Figure 3. Case studies by visualizing the attention of the last cross-modal attention layer in DeVLBert (the left for each case) and ViLBERT (right).

deconfounding, and architecture D-VLC by incorporating the cross-modal (inter-modality) deconfounding. The evaluation results are shown in Table 2, it can be seen that removing any deconfounding component will lead to a performance drop, which again verifies the effectiveness of the proposed framework.

Comparison with SOTA pretraining methods and task-specific downstream models. DeVLBert (with architecture D-VLC) achieves consistent performance improvement over task-specific SOTA models, including SCAN [3], BUTD [1], and over pretraining baselines, including ViLBERT [6], and VisualBERT [4]. DeVLBert achieves comparative performance with InterBert [5] that also incorporates in-domain dataset MSCOCO for pretraining.

3.2. Case Studies

We visualize the image region with the biggest attention weight for each language word (See Figure 3). The results indicate that: 1) **The attended visual tokens (object boxes) of DeVLBert are more accurate than those of ViLBERT.** By "accurate", we mean the attended tokens are more useful for determining whether this image is locally relevant to the query sentence, and better as reasoning cues given the question. We further compute the conditional probability of the answer given word `sitting`, which shows that DeVLBert can generate less frequent but more accurate answers. 2) **The results of DeVLBert yields less cognitive errors or spurious correlations**. For example, in case C_{11}, ViLBERT considers "person with wedding veil" as the "bride", and view the man as "bride" by mistake. The conditional probabilities under C_{22} and C_{31} show DeVLBert can learn to pay less attention to spuriously correlated tokens such as `sky` and `highway` by deconfounding.

4. Conclusion

In this paper, we propose to mitigate the spurious correlations for out-of-domain visio-linguistic pretraining. The fact that each output token is connected with all input tokens in Bert, and the pure association nature of masked token modeling objective makes the problem more severe. We borrow the idea of back-door adjustment to propose four novel Bert-style architectures as DeVLBert for out-of-domain pretraining. We conduct extensive quantitative evaluations as well as ablation studies to discuss the empirical effectiveness of different architectures.

References

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

[2] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *KDD*, 2018.

[3] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[4] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, 2019.

[5] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *CoRR*, 2020.

[6] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, page 13–23, 2019.

[7] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Taylor&Francis, 2020.

[8] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[9] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, 2020.

[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[11] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *ACM MM*, 2020.