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

Shengyu Zhang\textsuperscript{1}\textsuperscript{*}, Tan Jiang\textsuperscript{1}\textsuperscript{*}, Tan Wang\textsuperscript{2}, Kun Kuang\textsuperscript{1}\textsuperscript{†}, Zhou Zhao\textsuperscript{1}, Jianke Zhu\textsuperscript{1}, Jin Yu\textsuperscript{3}, Hongxia Yang\textsuperscript{3}\textsuperscript{†}, Fei Wu\textsuperscript{1}\textsuperscript{†}

\textsuperscript{1} Zhejiang University, \textsuperscript{2} University of Electronic Science and Technology of China, \textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{6}, part of their success can be traced back to the introduction of \textit{in-domain} pretraining datasets besides the Conceptual Caption\textsuperscript{8} dataset. By \textit{in-domain}, we refer to those datasets used in both pretraining and downstream tasks. However, out-of-domain pretrain-

\textsuperscript{1}Please refer to the full version of this paper\textsuperscript{11} for better clarity.

\textsuperscript{2}https://github.com/shengyuzhang/DeVLBert

\textsuperscript{*}These authors contributed equally to this work.

\textsuperscript{†}Corresponding Authors.

---

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{An illustration of the transition from traditional association-based learning to causal intervention-based learning.}
\end{figure}
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 deconfounded visio-linguistic representations. We name this kind of architectures as DeVL-Bert, which refers to Deconfounded Visio-Linguistic Bert. DeVL-Bert 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 DeVL-Bert architectures. Out-of-domain pretraining with three downstream vision-language tasks demonstrate that DeVL-Bert can boost the generalization ability by mitigating dataset biases.

2. Deconfounded Vxio-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 function can be re-written as:

$$P(Y|X) = \sum_z P(Y|X,z) P(z|X),$$

(1)

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

$$P(Y|do(X)) = \sum_z P(Y|X,z) P(z).$$

(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:

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

(3)

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

$$P(Y|do(X)) \approx \mathbb{E}_z [f_c(x,z)].$$

(4)

In this paper, we model the term $f_c(x,z)$ by the feedforward neural network $W_c[x, \alpha_y(z) \ast z]$, where $[, \cdot]$ denotes the concatenation operation and $\alpha_y(z)$ denotes the importance factor that is parameterized by $y$. Formally, we have:

$$\alpha_y(z) = \frac{(W_yy)^T(W_zy)}{\sum_{\nu \neq y} (W_y\nu)^T(W_\nu z)},$$

(5)

$$P(Y|do(X)) = \sigma(W_c[x, \sum_z P(z) \ast \alpha_y(z) \ast z]).$$

(6)

where $y/v$ is the feature representation of $Y/v$. $\zeta$ 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 $w_t$ as $x_t$ since $w_t$ 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 $y$ (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 $y_t$ while the final representations of all unmasked tokens can be viewed as $x_t$. This design is efficient without an extra inference process. The time complexity is $O(N_u \ast N_m)$, 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 $\{x_k\}_{k=1,\ldots,t-1,t+1,\ldots,N_u}$. 

**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:

$$\alpha_r(z) = \frac{(W_r r)^T(W_r z)}{\sum_{\nu \neq r} (W_r \nu)^T(W_\nu z)},$$

(7)

$$P(Y|do(X)) = \sigma(W_c \sum_z P(z) \ast \alpha_r(z) \ast z).$$

(8)

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

---

**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.
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 1. Comparison between DeVLBert and other competitors, including ViLBERT which only uses out-of-domain pretraining datasets, VisualBERT only uses in-domain datasets, and Inter-Bert using both.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Image Retrieval (IR)</th>
<th>Zero-shot IR</th>
<th>VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAN [3]</td>
<td>48.6</td>
<td>77.7</td>
<td>85.2</td>
</tr>
<tr>
<td>BUTD [1]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>*VisualBERT [4]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>*InterBert [5]</td>
<td>61.9</td>
<td>87.1</td>
<td>92.7</td>
</tr>
<tr>
<td>*ViLBERT [6]</td>
<td>38.2</td>
<td>84.9</td>
<td>91.5</td>
</tr>
<tr>
<td>DeVLBert</td>
<td>61.6</td>
<td>87.1</td>
<td>92.6</td>
</tr>
</tbody>
</table>

4.1. 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
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


3.2. Case Studies

We visualize the image region with the biggest attention produced by 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 generates less confident 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