# DeVLBert: Learning Deconfounded Visio-Linguistic Representations

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] College of Computer Science and Technology, Zhejiang University

[2] University of Electronic Science and Technology of China

[3] Alibaba Group

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

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

wangt97@hotmail.com

## ABSTRACT

In this paper, we propose to investigate the problem of 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, 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.

## CCS CONCEPTS

• **Computing methodologies → Transfer learning**.

## KEYWORDS

Multi-modal pretraining; Out-of-domain; Debias; Backdoor adjustment; Bert

**ACM Reference Format:**

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, Fei Wu. 2020. DeVLBert: Learning Deconfounded Visio-Linguistic Representations. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413518



*MM '20, October 12–16, 2020, Seattle, WA, USA*



ACM ISBN 978-1-4503-7988-5/20/10…$15.00

https://doi.org/10.1145/3394171.3413518

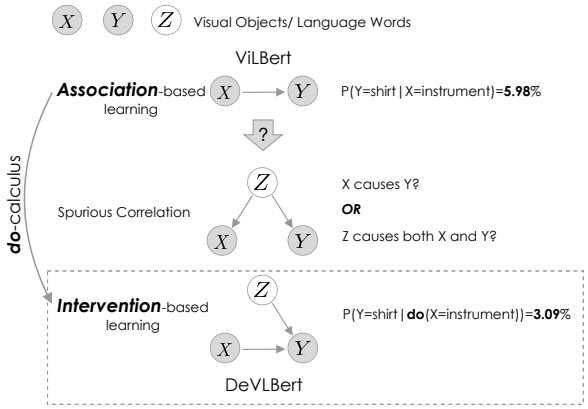


**Figure 1: An illustration of the transition from traditional association-based learning to causal intervention-based learning. The critical difference is that the intervention mitigates the spurious correlation by blocking the backdoor path $Z \rightarrow X$ and thus controlling the condition $X$.**

## 1 INTRODUCTION

Since early attempts that pretrain a backbone model [14, 18, 39] on large-scale dataset [10] and then transfer the knowledge to numerous computer vision tasks, pretraining has become a hallmark of the success of deep learning. More recently, the volume of transformer-based and Bert-style pretraining models [9, 11, 12, 27, 40] has grown tremendously in the research field of natural language processing and has achieved state-of-the-art performance in various NLP tasks. Likewise, the success of Bert-style pretraining techniques has been transferred to the research field of the intersection of vision and language [24, 29, 42–44].

Despite the significant progress that recent methods have made over the initiative work ViLBert [29], part of their success can be traced back to the introduction of *in-domain* pretraining datasets

[*] These authors contributed equally to this work.

[†] Corresponding Authors.

Work was performed when S. Zhang and T. Jiang were interns at Alibaba Group.

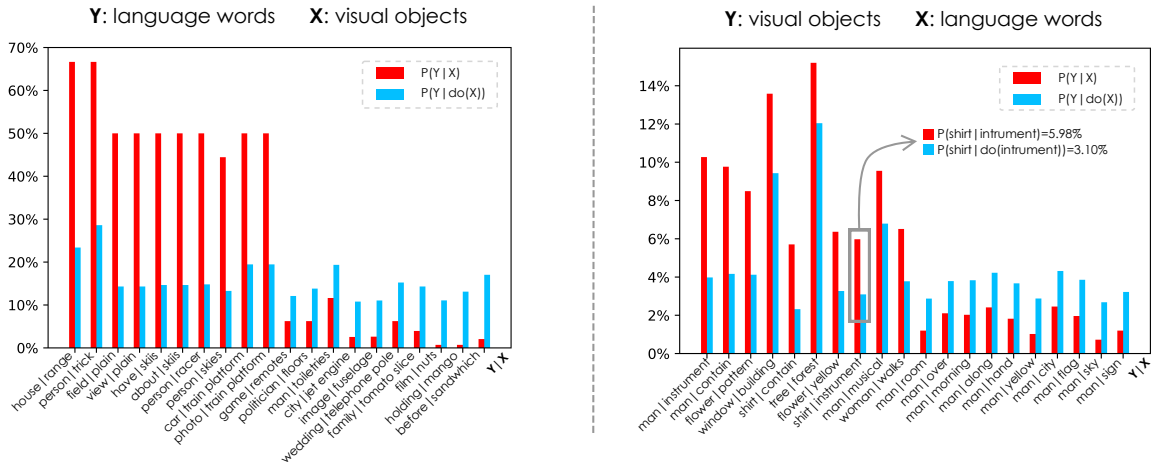


**Figure 2: The conditional probabilities of a token Y (visual object or language word) given another token X (from the other modality) before intervention (*i.e.*, P(Y|X), colored in red) and after intervention (*i.e.*, P(Y|do(X), colored in blue).**

besides the Conceptual Caption [37] dataset. By *in-domain*, we refer to those datasets used in both pretraining and downstream tasks, such as MSCOCO [26], and Visual Genome [17]. 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.

A fundamental requirement for out-of-domain transfer learning is to mitigate the biases from the pretraining data [49], which may be useful for the in-domain testing but harmful for out-of-domain testing [19] due to the *spurious correlation* [34]. To verify such existence of the correlation biases, we follow [49] to conduct a toy experiment on Conceptual Caption dataset. We observe that the conditional probability of `shirt` (visual object) given the `instrument` (language word) is large, *i.e.*, $p$(shirt|instrument) = 5.98%, but there are actually no robust relationships between them. 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 [20] and borrow the idea of the backdoor adjustment (also known as covariate adjustment or statistical adjustment) [2, 30, 49] to mitigate these biases. As shown in Figure 1, the traditional *association-based* learning fashion may lead to the spurious correlation between two tokens (visual objects or language words) by a common cause, *i.e.*, the confounder. By introducing backdoor adjustment (deconfounding), the original conditional probability of $p$(shirt|instrument) can be adjusted to 3.10% (nearly half) with a **do** operator. The essence of deconfounding is to control the condition (`instrument`) from being affected by other potential confounders when assessing the effect on the outcome (`shirt`) given the condition, *i.e.*, **intervention**. In this way, the pure *association*-based pretraining becomes to the causal *intervention*-based pretraining. We note that our goal is not performing theoretically causal inference but learning generic and de-biased visio-linguistic representations that can well generalize to downstream tasks with unknown data distributions.

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 token without considering whether there are spurious correlations (*e.g.*, `shirt` cannot cause `instrument`, and vice versa) or not. More formally, masked token modeling, abbreviated as MTM, models the conditional probability $P(Y|X)$ as the distribution of $Y$ when observing $X$. $Y$ is the masked token and $X$ denotes the context information. The spurious correlation occurs when $X$ and $Y$ are confounded by a common cause $Z$, as depicted in Figure 1. Our goal is to model the interventional operation $P(Y|do(X))$, meaning the distribution of $Y$ when controlling $X$ to mitigate the correlation bias as we introduced before. Real-world cases concerning the conditional probabilities and the corresponding intervention results from the Conceptual Captions dataset can be found in Figure 2. In this paper, we propose several intervention-based BERT architectures to help learn deconfouned visio-linguistic representations. We name this kind of architectures as **DeVLBert**, which refers to ***De**confounded **V**isio-**L**inguisitic **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. Pretraining is performed on the Conceptual Caption dataset which most downstream tasks are not built on, *i.e.*, out-of-domain dataset. We evaluate the effects of these architectures on three downstream cross-modal tasks, including text-to-image retrieval [48], zero-shot text-to-image retrieval, and visual question answering [3]. We also conduct case studies to evaluate DeVLBert from the human perspective, and demonstrate that mitigating dataset biases boosts the generalization ability.

The main contributions of our work are summarized as follows:

- We investigate the problem of out-of-domain pretraining, where pre-trained models are transferred to downstream tasks with unknown data distributions.
- We propose the novel DeVLBert framework designed for the Bert-style pretraining architectures with causal intervention

to mitigate the spurious correlations caused by the context-based proxy tasks.
- We devise four implementations of the DeVLBert framework [1] and discuss the empirical performance on several downstream tasks. The advantages of the DeVLBert are demonstrated by quantitative experiments, ablation studies, and case studies.

## 2 RELATED WORKS

### 2.1 Visio-linguistic Pretraining

Visio-linguistic (cross-modal) pretraining is a nascent research area that attracts considerable interests in recent years due to their strong ability of knowledge transfer. Existing works [8, 23, 24, 29, 42, 45] are mainly based on the Bert framework, including single-stream models and two-stream models. Besides the model structure, these methods differ mainly in the pretraining datasets and proxy tasks. Considerable methods incorporate in-domain datasets for pretraining, such as MSCOCO [26] and Visual Genome [17]. , which means the datasets are shared in the processes of both pretraining and downstream task training. In this paper, we investigate the particular out-of-domain pretraining problem, which can be a more general setting in the real-world.

### 2.2 Causality in Vision & Language

It is of increasing research interests in computer vision that attempts to borrow useful analysis tools from causality. Typical works are concerning object tracking [21, 52], interpretable Learning [15], image classification [7, 28], and image generation [5, 16]. More recently, Wang *et al.*[49] proposes the VC R-CNN framework for visual representation learning. They employ the causal intervention to deal with spurious correlation within datasets for visual common sense learning. However, VC R-CNN solely concerns intervention for visual domain. Alleviating the spurious correlations between vision and language in visio-linguistic pretraining can be also necessary, especially for cross-modal downstream tasks.

There are also considerable works that explore causality in natural language processing, varying among relation identification [31], text classification [50], and question answering [38]. More recently, some causality-related techniques also emerge in the research field of cross-modality. For example, visual dialogue [35], scene graph generation [46] and VQA [32]. Different from these works, DeVLBert investigates generic representation learning.

## 3 VISIO-LINGUISTIC BERT

### 3.1 Bidirectional Transformer

We choose Bidirectional Transformer (Bert) as our backbone structure, which can be the main-stream pretraining architecture for both natural language pretraining and visio-linguistic pretraining. For brevity, we take natural language pretraining as an example to illustrate the Bert structure. Given a language sequence $S = \{w_t\}_{t=1,\dots,N_w}$, Bert is pretrained to produce the contextualized word representations $\mathbf{S} = \{\mathbf{w}_t\}_{t=1,\dots,N_w}$. We note that the first token is often a pre-defined "[CLS]" and the learned representation for this token denotes the global representation of the whole sequence. Bert is composed of $N_l$ Transformer layers where each of the layers will output a feature sequence $\mathbf{S}^l = \{\mathbf{w}_t^l\}_{t=1,\dots,N_w}$. We view the feature sequence of the last layer as the final pretrained representations, *i.e.*, $\mathbf{w}_t = \mathbf{w}_t^{N_l}$. Each Transformer layer consists of a multi-head self attention module and a feed-forward module. Formally, such a process can be formulated as:

$$\mathbf{w}^l = \text{MultiHeadAttention}(\mathbf{w}^{l-1}), \tag{1}$$

$$\tilde{\mathbf{w}}^l = \text{LayerNorm}(\mathbf{w}^{l-1} + \mathbf{w}^l), \tag{2}$$

$$\hat{\mathbf{w}}^l = \text{FeedForward}(\tilde{\mathbf{w}}^l), \tag{3}$$

$$\mathbf{w}^l = \text{LayerNorm}(\tilde{\mathbf{w}}^l + \hat{\mathbf{w}}^l). \tag{4}$$

Each self attention head connect all pairs of input and output positions [47]:

$$\text{SelfAttention}\ (\mathbf{S}) = \text{softmax}\left(\frac{Q(\mathbf{S})K(\mathbf{S})^T}{\sqrt{D_w}}\right)V(\mathbf{S}), \tag{5}$$

while $Q$ and $K$ are learnable query transformation and key transformation to compute the attention weights. $V$ is the learnable linear transformation to obtain the value context. $D_w$ is the feature dimension of keys, values, and queries. This design yields a global correlation, *i.e.*, the final representation of each word token will be correlated with all other words. Since this modeling schema has no sense of word order (in the sequence) unlike RNNs, it is necessary add the position signals onto each word embedding. There are some techniques doing this, such as position embedding [13] and position encoding [47], and we follow the position embedding for simplicity. For the $t$th word token, the initial representation can be $\mathbf{w}_t^0 = \mathbf{w}_t^e + p_t$, where $\mathbf{w}_t^e$ denotes the word representation taken from the embedding layer and $p_t$ denotes the embedding for the position index $t$.

This structure can be easily transferred to the visual domain if the visual words are reasonably defined. A simple yet effective approach is to view the sub-regions of interest as visual words. More concretely, object detectors, such as Faster-RCNN [36], will be used to extract object bounding boxes and object feature maps. The original object representations $\mathbf{O}^e = \{\mathbf{o}_i^e\}_{i=1,\dots,N_v}$ are obtained by global average pooling over the object feature maps. Similar to the language pretraining, we add a global representation $\mathbf{o}_{[\text{G}]}^e = 1/N_v \sum_{i=1}^{N_v} \mathbf{o}_i^e$ at the beginning of the sequence. To be aware of the position signals of objects, we encode the information within bounding boxes to obtain position encodings. Concretely, each bounding box can be represented as a 5-d vector, including the normalized top-left coordinates, the normalized bottom-right coordinates, and the scaling factor. We note that for the global representation, the bounding box refers to that of the entire image. The position encoding vector of the same dimension as the object representation is then obtained by a feed-forward network.

### 3.2 Two-stream Visio-Linguistic Modeling

Existing Bert-like visio-linguistic pretraining methods can be roughly categorized into single-stream architectures and two-stream architectures. We are following the two-stream architectures [29, 42, 45] since they keep the independence of each modality as well as modeling the interaction across different modalities. The only difference

[1] Code will be released at https://github.com/shengyuzhang/DeVLBert

between the visio-linguistic Transformer layer and the modality-specific Transformer layer lies in the queries. Formally, the attention head in visio-linguistic Transformer layer for the language side can be formulated as:

$$\text{SelfAttention}\,(\mathbf{S},\mathbf{O}) = \text{softmax}\left(\frac{Q(\mathbf{S})K(\mathbf{O})^T}{\sqrt{D_w}}\right)V(\mathbf{O}), \tag{6}$$

Intuitively, this process is designed to borrow language-related information (language features **S** as queries) from the visual features (**O** as keys and values). Likewise, for the visual side, we have:

$$\text{SelfAttention}\,(\mathbf{O},\mathbf{S}) = \text{softmax}\left(\frac{Q(\mathbf{O})K(\mathbf{S})^T}{\sqrt{D_o}}\right)V(\mathbf{S}). \tag{7}$$

where $D_o$ denotes the feature dimension of visual queries. By combining the visio-linguistic Transformer layer and the modality-specific Transformer layer and further stacking the combined layers, we obtain the main structure of two-stream visio-linguistic Bert.

### 3.3 Pretraining Proxy Tasks

Masked language modeling (MLM) and masked object modeling (MOM) are popular context-based proxy tasks for language and vision, respectively. We group MLM and MOM into masked token modeling (MTM) in this paper for brevity. As the name (masked) implies, MTM often randomly masks tokens in a sequence by a given probability and replace them with special tokens, such as "[MASK]". Then, Bert is required to make predictions to recover these tokens. Typical recovering strategies include token vocabulary classification and feature regression. For MLM, we follow the standard practice in original BERT[11]. For MOM, 15% proportion of the objects will be masked ready for recovering. Since no objects are masked during testing, which yields a setting gap between training and testing, we replace 10% of the masked objects with their original representations. We note that these objects are still required to be *recovered*, which are different from the initially unmasked objects. By recovering, we mean the object classification with soft labels, which come from the object detectors.

Visio-linguistic alignment follows the original design of the next sentence prediction objective in natural language pretraining. By the probability of 50%, the originally paired sentence will be replaced by a random sampled unpaired sentence. The Bert model is required to predict whether the language sequence and the objects sequence are aligned. We employ a simple feed-forward neural network to compute the alignment score based on the global representation of words/objects sequence, *i.e.*, $\mathbf{w}_{[CLS]}$ and $\mathbf{o}_{[G]}$, where $\mathbf{o}_{[G]}$ denotes the final representation of the special global object $o_{[G]}$.

## 4 DECONFOUNDED VSIO-LINGUISTIC BERT

### 4.1 Bert in the causal view

As illustrated in section 3.1, 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 $Y$. If such confounders are not controlled in modeling, some false conclusion about $X$ and $Y$ may be drawn as part or all of the effect might come from $Z$. The key to alleviate the spurious correlations is to control the confounders when evaluating the causal effect of $Y$ given $X$, *i.e.*, intervention-based modeling $P(Y|do(X))$. Our framework borrows the idea of backdoor adjustment [30, 49]. Formally, by the Bayes Rule, the conventional likelihood can be re-written as:

$$P(Y|X) = \sum_z P(Y,z|X) = \sum_z P(Y|X,z)\underline{P(z|X)}, \tag{8}$$

By using the *do*-calculus, we remove any incoming influence to the intervened variable, *i.e.*, $X$. By the definition of *do*-calculus, we have:

$$P(Y|do(X)) = \sum_z P(Y,z|do(X)) \tag{9}$$

$$= \sum_z P(Y|do(X),z)P(z|do(X)) \tag{10}$$

$$= \sum_z P(Y|X,z)\underline{P(z)}. \tag{11}$$

The proof of transitions $P(Y|do(X)) = P(Y|X,z)$ and $P(z|do(X) = P(z)$ can be found in the book [4] and in several following works [35, 49]. The prior probability of each $z$ can be easily pre-counted before training following [49]. 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 [41, 51] to approximate the expensive sampling and separate modelling as [49] does. Formally, if the last objective is classification, we can re-write the following terms:

$$P(Y|X,z) = \text{softmax}\left(f_c(\mathbf{x},\mathbf{z})\right), \tag{12}$$

$$P(Y|do(X)) = \mathbb{E}_z\left[\text{softmax}\left(f_c(\mathbf{x},\mathbf{z})\right)\right], \tag{13}$$

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

$$\mathbb{E}_z\left[\text{softmax}\left(f_c(\mathbf{x},\mathbf{z})\right)\right] \overset{NWGM}{\approx} \text{softmax}\left(\mathbb{E}_z\left[f_c(\mathbf{x},\mathbf{z})\right]\right). \tag{14}$$

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 $\mathbf{y}$. The introduction of y-dependent confounder importance re-weighting strategy follows VC R-CNN [49]. This formulation is reasonable in the sense that a particular $z$ that well correlates with $Y$ have high probability to be the confounder of $Y$ and $X$. Formally, we have:

$$\alpha_y(\mathbf{z}) = \frac{(\mathbf{W}_y\mathbf{y})^T(\mathbf{W}_z\mathbf{z})}{\sum_{v\neq\varsigma}(\mathbf{W}_y\mathbf{y})^T(\mathbf{W}_z\mathbf{v})}, \tag{15}$$

$$P(Y|do(X)) = \text{softmax}\left(\mathbf{W}_c\left[\mathbf{x}, \sum_z P(\mathbf{z}) * \alpha_y(\mathbf{z}) * \mathbf{z}\right]\right). \tag{16}$$

where $\mathbf{y}/\mathbf{v}$ is the feature representation of $Y/v$. $\varsigma$ denotes the confounder that has the same token class as $Y$. For example, if the predicted token is $y = \texttt{cat}$, it is unreasonable to take $z = \texttt{cat}$ as a potential confounder for predicting $y = \texttt{cat}$. We note that the corresponding weight $\alpha_y(\varsigma)$ is thus 0. Now the problem of how to perform intervention-based learning is transformed to define how Bert models the feature representation of $X$ and $Y$. In this paper,

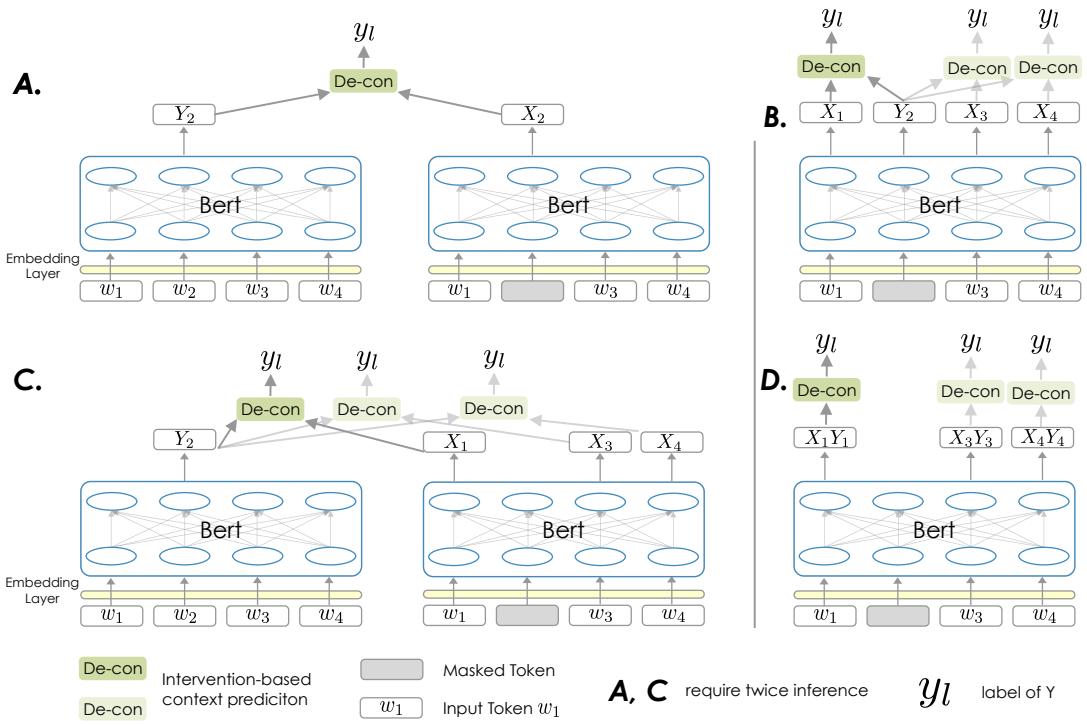


**Figure 3: A vivid illustration of four intervention formulations for Bert-style training. Deciding the forms of X and Y in Bert is essential for further intervention-based context prediction. Design A&C require twice inference. Design A&B replace the frequently used masked-token-modeling (MTM) objective while Design C&D are independent of the MTM.**

we propose several implementations. We note that in the following illustrations, each case is concerning only one masked token, and it is easy to extend the framework to the general case with multiple masked tokens.

- **Design A**. We firstly investigate how to harness masked token modeling with intervention, since 1) MTM is among the most popular pretraining proxy tasks. 2) MTM is solely based on likelihood estimation, which might introduce spurious correlations. 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). However, it is not easy to find $\mathbf{y}$ which contains the information of word $w_t$ in single inference. We choose to run another inference with no masked tokens. In this way, the final representation of word $w_t$ can be viewed as $\mathbf{y}$. This implementation is depicted in Figure 3 A.
- **Design B**. Figure 3 B depicts another design to harness MTM. Under the framework of MTM, Bert leverages and aggregates $w_t$-related information from the context to predict the label of $w_t$. In this perspective, 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 $\{\mathbf{x}_k\}_{k=1,\ldots,t-1,t+1,\ldots,N_w}$. 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 3 C, Design C is a variant of Design A and views the final representations of all unmasked tokens as $\{\mathbf{x}_k\}_{k=1,\ldots,t-1,t+1,\ldots,N_w}$.
- **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. By non-intrusive, we mean that this design makes fewer modifications, *i.e.*, without another forward run and without hurting the original MTM objective. With the time complexity of $O(N_u)$, Design D is more efficient than the double-run designs as well as Design B. We note that in this design, the modeling of $P(Y|do(X))$ is slightly different:

$$\alpha_r(\mathbf{z}) = \frac{(\mathbf{W}_r\mathbf{r})^T(\mathbf{W}_z\mathbf{z})}{\sum_z(\mathbf{W}_r\mathbf{r})^T(\mathbf{W}_z\mathbf{z})}, \tag{17}$$

$$P(Y|do(X)) = \text{softmax}\left(\mathbf{W}_c \sum_z P(\mathbf{z}) * \alpha_r(\mathbf{z}) * \mathbf{z}\right). \tag{18}$$

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 the concatenation operation. Using both $\mathbf{y}$ and $\mathbf{x}$ to re-weight the importance of each $\mathbf{z}$ is also reasonable in the sense that a particular $z$ that well correlates with both $Y$ and $X$ have high probability to be the confounder of $Y$ and $X$.

## 4.2 Intra- & Inter-modality Intervention

**Vision deconfounding & Vision Confounder Set.** It is infeasible to take each particular object (in some image) as a potential confounder as there can be numerous objects in pretraining datasets. Following VC R-CNN [49], 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). The size of the confounder set (1,600) is equivalent to the number of object classes that are pre-defined by the pre-trained object detector. 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.** Likewise, it is infeasible to take each particular word token (in some sentence) as a potential confounder, and it is also expensive to discuss all high-level words since there are about 30,000 words in the Bert vocabulary. In this paper, we choose nouns as potential confounders since 1) nouns are content words that have meaning or semantic value [53]; 2) the role of nouns is similar to the role of objects in image, which might ease the inter-modality intervention. Specifically, we use the NLTK toolkit [6] to perform Part-of-Speech Tagging, and choose word tokens of which the tags belong to ["*NN*", "*NNS*", "*NNP*", "*NNPS*"] as potential confounders. There are in total 156 potential confounders in the language confounder set (after removing nouns with low-frequencies since such words have less chance to be 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. Similarly, for language deconfounding, $Y$ and $X$ are only selected from the final representations of the language words, and confounders in the language confounder set are discussed. We note that only noun words are considered as $X$ and $Y$ since they are of high probability to pose spurious correlations with visual objects. For Design A and B, the MLM objective is replaced by the intervention objective for masked noun words, and others words still have chance to process masked prediction. For Design C and D, the masked prediction objective is not affected by the intervention objective.

**Inter-modality Intervention.** It is necessary to perform inter-modality (or cross-modal) intervention since with the two-stream visio-linguistic modeling, the token representations of each modality contain information from other modalities, which may lead to spurious correlations without inter-modality intervention. We observe that in the Conceptual Caption dataset, the conditional probability of visual object "`shirt`" give the word "`instrument`" is about 6% while `shirt` and `instrument` have no causal relationship but might have a common cause, *i.e.*, the visual object or the language word `person`. Specifically, 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.

We note that both the MTM and the intervention-based objective conduct object classification for vision, and word prediction for language. The difference between intervention and MTM is that intervention discusses the effect of $Y$ given $X$ and each potential confounder $z$, which helps mitigate the spurious correlations.

# 5 EXPERIMENTS

## 5.1 Experiment Setup

**Pretraining DeVLBert.** We follow ViLBERT [29] to pretrain DeVLBert on the Conceptual Caption [37] dataset, which is an out-of-domain dataset that has little data overlap with most downstream tasks. Images and raw descriptions are harvested from HTML pages that contain images and Alt-text attributes. Then, automatic language cleaning pipelines are developed to obtain the final image captions that are clean, informative, but less similar to the human-annotated captions in datasets of downstream tasks. In other words, the Conceptual Caption dataset serves as an excellent dataset for out-of-domain pretraining. Due to broken or expired links by the time we downloaded, we use around 3.04 million <image, caption> pairs for pretraining, which is smaller than the original 3.3 million dataset when first published, and also smaller than the 3.1 million dataset used in ViLBERT. To make our results comparable to the previous out-of-domain pretraining work, *i.e.*, ViLBERT, we are following the exact pipeline as theirs, including the initialization of the linguistic stream and visual region feature extraction.

**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[29]. We note that our goal is not achieving the state-of-the-art with bells&whistles but demonstrating the effectiveness of mitigating spurious correlations of DeVLBert for out-of-domain pretraining. Besides the quantitative evaluations, we conduct user studies that qualitatively show whether and how DeVLBert achieves better results by mitigating biases.

**Hardware & Software Configuration** We implement the models in python3.6 and PyTorch 1.1.0 [33], and train the models on a Linux server equipped with 8 NVIDIA V100-SXM2-16GB GPUs.

## 5.2 Quantitative Evaluation

By quantitative evaluation, we care about a few issues listed below:

**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:

- Most of the architectures obtain performance gain on at least one of the tasks, which demonstrates the effectiveness of intervention-based learning.
- The twice inference design achieves inferior results on the zero-shot image retrieval task. Partially due to the complexities introduced by another inference, it might take more iterations to converge, which can be expensive. Moreover,

**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 [22] | 48.6 | 77.7 | 85.2 | - | - | - | - | - |
| BUTD [1] | - | - | - | - | - | - | 65.3 | 65.7 |
| $^{\bullet}$VisualBERT [24] | - | - | - | - | - | - | 70.8 | 71.0 |
| $^{\odot}$InterBert [25] | 61.9 | 87.1 | 92.7 | 49.2 | 77.6 | 86.0 | 70.3 | 70.6 |
| $^{\circ}$ViLBERT [29] (Baseline) | 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 |

**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** |

similar to the results shown in VC R-CNN [49] that the performance of directly using the pre-trained commonsense features (intervention-based) is lower than that of the original features (association-based) while combining these two features would achieve the best performance. In our case, combining the pretrained deconfounding features with the knowledge in the downstream task (regular image retrieval) achieves better results in zero-shot IR.

- 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.
- 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.
- 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.

**Table 3: Comparisons between DeVLBert and VC R-CNN (with Bert as language feature extractor).**

| Method | Image Retrieval (IR) | | | VQA | |
|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | test-dev | test-std |
| VC R-CNN [49] | 15.0 | 40.4 | 54.7 | 53.2 | 53.6 |
| DeVLBert | 39.2 | 70.2 | 80.5 | 53.5 | 53.9 |

**Do both intra-modality intervention and 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 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.

**How does DeVLBert perform comparing with existing visio-linguistic pretraining methods and task-specific downstream SOTA models?** We view the architecture D-VLC as DeVLBert due to its empirical effectiveness. As shown in Table 1, when compared to task-specific SOTA models, including SCAN [22] for image retrieval, BUTD [1] for visual question answering, DeVLBert yields a large margin improvement over these methods. More importantly, while the baseline method, *i.e.*, ViLBERT [29], cannot beat in-domain pretraining methods, including VisualBERT [24] using MSCOCO as the pretraining dataset, and InterBert using both Conceptual Caption and MSCOCO as pretraining datasets, DeVLBert obtains a performance boost over the VisualBert and InterBert on the VQA task and achieves comparable results to the InterBert on the image retrieval task. On zero-shot image retrieval, DeVLBert cannot beat InterBert, which is a reasonable result since MSCOCO is the testing dataset for zero-shot image retrieval, and InterBert uses MSCOCO for pretraining.

**How does DeVLBert perform comparing with VC R-CNN?** The comparison results with VC R-CNN [49], which learns visual commonsense features by intervention, are listed in Table 3. We use VC R-CNN as the visual feature extractor and vanilla Bert as the language feature extractor. For fair comparison, we do not fine-tune DeVLBert on the downstream datasets either, and view DeVLBert as visio-language feature extractor. We mean-pool the extracted features for each modality and concatenate the pooled features from two modalities. We use two layer MLP as classifier with hidden size twice as large as the input feature size. According

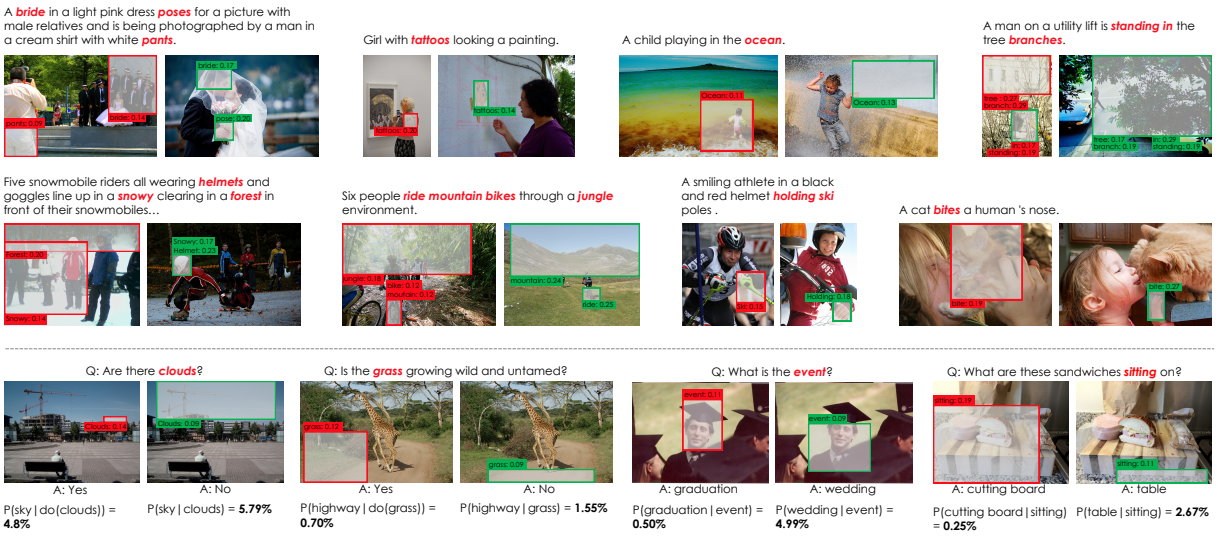


**Figure 4: Case studies by visualizing the attention of the last cross-modal attention layer in DeVLBert (the left for each case) and ViLBERT (right). Cases are sampled from the testing/validation set of downstream tasks, *i.e.*, image retrieval (top), and VQA (bottom). The labeled word is the attention query word from the input sentence, and the number is the corresponding attention weight.**

to Table 3, DeVLBert achieves large performance gain over VC R-CNN on multi-modal matching tasks (Image Retrieval) and competitive results on VQA, which shows cross-modal pretraining and deconfounding are essential for cross-modal downstream tasks.

## 5.3 Case Studies

To further evaluate the effectiveness of DeVLBert from the human perspective, we follow [49] to conduct case studies on the testing/validation set of downstream tasks, including image retrieval and VQA (See Figure 4). For image retrieval, given a query sentence, we select the top answer image of DeVLBert (left) and ViLBERT (right). Compared to the VC R-CNN that focusing on visual attention, here we are especially interested in cross-modal attention, which is essential for visual-language tasks. Concretely, there are multiple co-attention blocks in both ViLBERT and DeVLBert, and we select the last block to obtain task-oriented association (the closer to the classification layer, the more task-oriented). Each co-attention block still contains multiple attention heads, we take the average attention map of all heads for visualization. We select the box with the biggest attention weight for each word. 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. For example, in the case $C_{34}$, which denotes the case in the 3rd row and the 4th column, the attended box of ViLBERT (right) directly focuses on `sitting` and fails to consider the `sandwiches sitting`, *i.e.*, question-specific context, while the attended box of DeVLBert is more accurate. 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. In case $C_{32}$, there is a spurious correlation between the word `grass` and the visual object `highway`, which drives the ViLBERT to attend to the region with both `grass` and `highway`. With such attended region, ViLBERT fail to realize that the grass is growing wild and untamed. The conditional probabilities under $C_{31}$ and $C_{32}$ show DeVLBert can learn to pay less attention to spuriously correlated tokens such as `sky` and `highway` by deconfounding.

## 6 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. The results show that DeVLBert can achieve promising numerical results compared to the baseline and even some in-domain visio-linguistic pretraining methods.

## 7 ACKNOWLEDGMENTS

The work is supported by the NSFC (61625107, 61751209, 61836002), National Key R&D Program of China (No. 2018AAA0101900, No. 2018AAA0100603), Zhejiang Natural Science Foundation (LR19F020006), Fundamental Research Funds for the Central Universities (2020QNA5024), and a research fund supported by Alibaba.

# REFERENCES


[1] 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.. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 6077–6086. http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html

[2] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering.. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 2425–2433. https://doi.org/10.1109/ICCV.2015.279

[4] Peter M Aronow and Fredrik Sävje. 2020. *The Book of Why: The New Science of Cause and Effect: Judea Pearl and Dana Mackenzie. New York: Basic Books, 2018, x+ 418 pp., ISBN: 978-0-46-509760-9*. Taylor & Francis.

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. 2019. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks.. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=Hyg_X2C5FX

[6] Steven Bird. 2006. NLTK: The Natural Language Toolkit.. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. https://www.aclweb.org/anthology/P06-4018/

[7] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. 2015. Visual Causal Feature Learning.. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*. 181–190. http://auai.org/uai2015/proceedings/papers/109.pdf

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning UNiversal Image-TExt Representations. *CoRR* abs/1909.11740 (2019). http://arxiv.org/abs/1909.11740

[9] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining.. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. 7057–7067. http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database.. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

[12] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation.. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. 13042–13054.

[13] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning.. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. 1243–1252. http://proceedings.mlr.press/v70/gehring17a.html

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition.. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[15] Jinkyu Kim and John F. Canny. 2017. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention.. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2961–2969. https://doi.org/10.1109/ICCV.2017.320

[16] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training.. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. https://openreview.net/forum?id=BJE-4xW0W

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, and et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks.. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. 1106–1114.

[19] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. 2018. Stable Prediction across Unknown Environments.. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. 1617–1626. https://doi.org/10.1145/3219819.3220082

[20] Li Lian Geng Zhi Xu Lei Zhang Kun Liao Beishui Huang Huaxin Ding Peng Miao Wang Jiang Zhichao Kuang, Kun. 2020. Causal Inference. *Engineering* (2020).

[21] Karel Lebeda, Simon Hadfield, and Richard Bowden. 2015. Exploring Causal Relationships in Visual Object Tracking.. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 3065–3073. https://doi.org/10.1109/ICCV.2015.351

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching.. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*. 212–228. https://doi.org/10.1007/978-3-030-01225-0_13

[23] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. *CoRR* abs/1908.06066 (2019). http://arxiv.org/abs/1908.06066

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR* abs/1908.03557 (2019). http://arxiv.org/abs/1908.03557

[25] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. InterBERT: Vision-and-Language Interaction for Multi-modal Pretraining. *CoRR* abs/2003.13198 (2020). https://arxiv.org/abs/2003.13198

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context.. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692

[28] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. 2017. Discovering Causal Signals in Images.. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 58–66. https://doi.org/10.1109/CVPR.2017.14

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. 13–23.

[30] Leland Gerson Neuberg. 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory* 19, 4 (2003), 675–685.

[31] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint Reasoning for Temporal and Causal Relations.. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 2278–2288. https://www.aclweb.org/anthology/P18-1212/

[32] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2020. Counterfactual VQA: A Cause-Effect Look at Language Bias. *arXiv preprint arXiv:2006.04315* (2020).

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

[34] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[35] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2019. Two Causal Principles for Improving Visual Dialog. *CoRR* abs/1911.10496 (2019). http://arxiv.org/abs/1911.10496

[36] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. 91–99.

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning.. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 2556–2565. https://www.aclweb.org/anthology/P18-1238/

[38] Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating Causal Embeddings for Question Answering with Minimal Supervision.. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 138–148. https://doi.org/10.18653/v1/d16-1014

[39] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition.. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1409.1556

[40] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation.. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. 5926–5936. http://proceedings.mlr.press/v97/song19d.html

[41] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958.

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations.. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=SygXPaEYvH

[43] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Contrastive Bidirectional Transformer for Temporal Representation Learning. *CoRR* abs/1906.05743 (2019). http://arxiv.org/abs/1906.05743

[44] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning.. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 7463–7472. https://doi.org/10.1109/ICCV.2019.00756

[45] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers.. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. 5099–5110. https://doi.org/10.18653/v1/D19-1514

[46] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased Scene Graph Generation from Biased Training. *CoRR* abs/2002.11949 (2020). https://arxiv.org/abs/2002.11949

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need.. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 5998–6008.

[48] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings.. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 5005–5013. https://doi.org/10.1109/CVPR.2016.541

[49] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual Commonsense R-CNN. *CoRR* abs/2002.12204 (2020). https://arxiv.org/abs/2002.12204

[50] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of Using Text Classifiers for Causal Inference.. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. 4586–4598. https://www.aclweb.org/anthology/D18-1488/

[51] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 2048–2057. http://proceedings.mlr.press/v37/xuc15.html

[52] Yuanlu Xu, Lei Qin, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. 2018. A Causal And-Or Graph Model for Visibility Fluent Reasoning in Tracking Interacting Objects.. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 2178–2187. http://openaccess.thecvf.com/content_cvpr_2018/html/Xu_A_Causal_And-Or_CVPR_2018_paper.html

[53] Zekun Yang and Tianlin Liu. 2020. Causally Denoise Word Embeddings Using Half-Sibling Regression.. In *Association for the Advancement of Artificial Intelligence*.