Causal Attention for Vision-Language Tasks

Xu Yang[1], Hanwang Zhang[1], Guojun Qi[2], Jianfei Cai[3]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore,
[2]Futurewei Technologies
[3]Faculty of Information Technology, Monash University, Australia,
s170018@e.ntu.edu.sg, hanwangzhang@ntu.edu.sg, guojunq@gmail.com, Jianfei.Cai@monash.edu

Abstract

We present a novel attention mechanism: Causal Attention (CATT), to remove the ever-elusive confounding effect in existing attention-based vision-language models. This effect causes harmful bias that misleads the attention module to focus on the spurious correlations in training data, damaging the model generalization. As the confounder is unobserved in general, we use the front-door adjustment to realize the causal intervention, which does not require any knowledge on the confounder. Specifically, CATT is implemented as a combination of 1) In-Sample Attention (IS-ATT) and 2) Cross-Sample Attention (CS-ATT), where the latter forcibly brings other samples into every IS-ATT, mimicking the causal intervention. CATT abides by the Q-K-V convention and hence can replace any attention module such as top-down attention and self-attention in Transformers. CATT improves various popular attention-based vision-language models by considerable margins. In particular, we show that CATT has great potential in large-scale pre-training, e.g., it can promote the lighter LXMERT [57], which uses fewer data and less computational power, comparable to the heavier UNITER [14]. Code is published in https://github.com/yangxuntu/lxmertcatt.

1. Introduction

Stemming from the strong cognitive evidences in selective signal processing [59, 50], the attention mechanism has arguably become the most indispensable module in vision and language models [66, 5, 3, 16, 11, 36]. Although its idiosyncratic formulation varies from task to task, its nature can be summarized as the following common Q-K-V notation: given a *query* $\mathbf{q}$, the attention mechanism associates $\mathbf{q}$ to each feature *value* $\mathbf{v}_i$ by using the normalized attentive weight $\alpha_i \propto \mathbf{q}^T\mathbf{k}_i$, where $\mathbf{k}_i$ is the *key* function of the value; thus, the resultant selective feature value — attention — is $\sum_i \alpha_i \mathbf{v}_i$. In a modern view, the attention can be understood as a feature *transformer* that encodes input query $\mathbf{q}$ by using the given values $\boldsymbol{V} = \{\mathbf{v}_i\}$ [60].

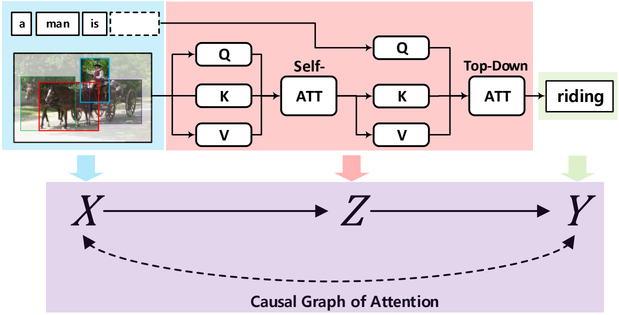

Figure 1. Top: an example of image captioner with a self-attention and a top-down attention modules. Bottom: the corresponding causal graph. The reason why the prediction is “riding” but not “driving” is explained in Figure 3.

Taking image captioning as an example in Figure 1, if $\mathbf{q}$ and $\boldsymbol{V}$ are both encoded from the input X, *e.g.*, the RoI features of an image, we call it *self-attention*; if $\mathbf{q}$ is changed to the sentence context, we call it *top-down attention*. Intuitively, self-attention is usually viewed as a non-local [65] (or graph [7]) convolution network that enriches each local value with global relationship features; top down-attention is used to enrich the context with the cross-domain relationship features [3]. Both of them can be combined and stacked into deep networks, serving as powerful multi-modal encoder-decoder transformer networks [32, 12].

As a bridge connecting the input feature X and the output label Y, the quality of attention — how reasonable the attentive weight α is — plays a crucial role for the overall performance. However, due to the fact that the attention weights are unsupervised, *e.g.*, there is no word-region grounding for the top-down attention or relationship dependency annotation for the self-attention, the weights will be inevitably misled by the dataset bias. For example, as shown in Figure 1, since there are many images captioned with “person riding horse” in the training data, self-attention learns to infer “riding” by building the dependency between “person” and “horse”. Then, given a test

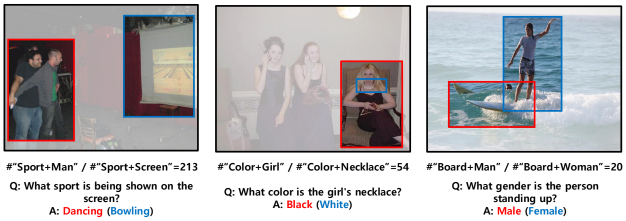
#"Sport+Man" / #"Sport+Screen"=213
Q: What sport is being shown on the screen?
A: Dancing (Bowling)
#"Color+Girl" / #"Color+Necklace"=54
Q: What color is the girl's necklace?
A: Black (White)
#"Board+Man" / #"Board+Woman"=20
Q: What gender is the person standing up?
A: Male (Female)

Figure 2. Before pre-training (*e.g.*, LXMERT [57]), attentions are correct (blue). After pre-training, attentions are wrong (red). This is because the co-occurrences of some concepts appear much more often than others, *e.g.*, "Sport+Man" appears 213 times more than "Sport+Screen" in the pre-training data.

image with "person driving carriage", this self-attention still tends to relate "person" with "horse" to infer "riding", but ignoring the "carriage". Unfortunately, such bias cannot be mitigated by simply enlarging the dataset scale, as most of the bias abides by the data nature — Zipf's law [48] and social conventions [19] — there are indeed more "red apple" than "green apple" or "person standing" than "person dancing". Therefore, as shown in Figure 2, large-scale pre-training may lead to even worse attentions.

The dataset bias is essentially caused by the confounder, a common cause that makes X and Y correlated even if X and Y have no direct causation. We illustrate this crucial idea in Figure 3. Suppose that the confounder C is the common sense[1] "person can ride horse", $C \to X$ denotes that a visual scene is generated by such knowledge, *e.g.*, the dataset curator observes and captures the common sense; $X \to M$ denotes the fact that the objects $M = \{\text{person}, \text{horse}\}$ can be detected (*e.g.*, Faster R-CNN [49]), whose object inventory is determined by $C \to M$; $M \to Y$ denotes the language generation for "person riding horse". Note that besides the legitimate causal path from image X via object M to Y, the "backdoor" path $X \leftarrow C \to M \to Y$ also contributes an effect to Y. Therefore, if we only train the model based on the correlation $P(Y|X)$ without knowing the confounding effect, no matter how large the amount of training data is, the model can never identify the true causal effect from X to Y [41, 52]. For example, if the confounder distribution varies from training to testing, *e.g.*, the common sense "person can ride horse" is dominantly more often than the common sense "person can drive carriage" in training, but the latter is more often than the former in testing, then $P(Y|X)$ based on "person can ride horse" in training will be no longer applicable in testing [42].

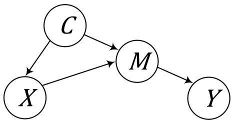

Figure 3. This expands the causal links of the confounding path $X \dashleftarrow\dashrightarrow Y$ in Figure 1
.

In this paper, we propose a novel attention mechanism called: *Causal Attention* (CATT), which can help the models identify the causal effect between X and Y, and thus mitigates the bias caused by confounders. It is based on the *front-door adjustment* principle that does not require the assumption of any observed confounder [40], and thus CATT can be applied in any domain where the attention resides. In this way, CATT is fundamentally different from existing deconfounding methods based on the backdoor adjustment [76, 64], which has to be domain-specific to comply with the observed-confounder assumption. Specifically, we first show that the conventional attention is indeed an improper approximation of the front-door principle, and then we show what is a proper one, which underpins CATT theoretically (Section 3.1).

We build CATT on the proposed In-Sample attention (IS-ATT) and Cross-Sample attention (CS-ATT), which abides by the Q-K-V operations (Section 3.2). In particular, the parameters of the Q-K-V operations can also be shared between both IS-ATT and CS-ATT to further improve the efficiency in some architectures. We replace the the conventional attention with CATT in various vision-language models to validate its effectiveness, including the classic Bottom-Up Top-Down LSTM [3], Transformer [60], and a large-scale vision-language pre-training (VLP) model LXMERT [57]. The experimental results demonstrate that our CATT can achieve consistent improvements for all of them. Significantly, our light LXMERT+CATT outperforms the heavy UNITER [14] on VQA2.0, *i.e.*, 73.04% vs. 72.91% on test-std split, and NLVR2, *i.e.*, 76.0% vs. 75.80% on test-P split, while we require much fewer pre-training burdens: 624 vs. 882 V100 GPU hours. Such comparisons show that our CATT has great potential in vision-language pre-training (VLP) tasks.

2. Related Work

Attention Mechanism. Conventional top-down attentions generally include the classic single-guidance fashion [5, 66, 69, 71] and the co-guidance fashion [34, 73]. They can be summarized as the query, key, value (Q-K-V) operation that also generalizes to self-attention [60, 65], which even be applied in pure vision tasks such as visual recognition and generation [11, 12]. As the attention weight is unsupervised, it is easily misled by the confounders hidden in the dataset. We exploit the causal inference to propose a novel CATT module to mitigate the confounding effect [44, 41]. As our proposed CATT complies with the Q-K-V convention, it has great potential in any model that uses attention.

Vision-Language Pre-Training. Inspired by the success of large-scale pre-training for language modeling [16, 47], researchers have developed some multi-modal Transformer-based Vision-Language Pre-training (VLP) models to learn task-agnostic visiolinguistic representations [32, 57, 29, 14, 78, 30, 28]. To discover the visiolinguistic relations across

[1] It is also well-known as the disentangled causal mechanism [56].

domains, a huge amount of data [53, 13, 26] are required for VLP. However, just as the language pre-training models tend to learn or even amplify the dataset bias [27, 37], these VLP models may also overplay the spurious correlation. We use the proposed CATT to help VLP models confront the bias.

Causal Inference. Causality [41, 52] provides researchers new methodologies to design robust measurements [56], discover hidden causal structures [9], generate counterfactual samples [58, 1, 25, 74], and confront various biases [62, 75, 76, 38, 20, 45]. These bias removal methods usually assume that the confounder is observable [75, 76] or domain-specific [19, 10]. In general, the confounder is unobservable and elusive. Compared with them, we exploit the front-door adjustment [40] with no observed-confounder assumption to mitigate the dataset bias. To tackle the sampling challenge in the front-door adjustment, we propose two effective approximations called In-Sample Sampling and Cross-Sample Sampling.

3. Causal Attention

3.1. Attention in the Front-Door Causal Graph

We retrospect the attention mechanism in a front-door causal graph [44, 41] as shown in the bottom part of Figure 1, where the causal effect is passed from the input set X to the target Y through a mediator Z. By this graph, we can split the attention mechanism into two parts: a selector which selects suitable knowledge Z from X and a predictor which exploits Z to predict Y. Take VQA as the example, X is a multi-modality set containing an image and a question, then the attention system will choose a few regions from the image based on the question to predict the answer. We usually use the observational correlation $P(Y|X)$ as the target to train an attention-based model:

$$P(Y|X) = \underbrace{\sum\nolimits_z P(Z=z|X)}_{\text{IS-Sampling}} P(Y|Z=z), \tag{1}$$

where z denotes the selected knowledge and **IS-Sampling** denotes In-Sample sampling since z comes from the current input sample X.

However, as discussed in Introduction, since the selection is an unsupervised process, the predictor may be misled by the dataset bias when training it by Eq. (1). In causal terms, this means that the predictor may learn the spurious correlation brought by the backdoor path $Z \leftarrow X \leftrightarrow Y$[1] instead of the true causal effect $Z \rightarrow Y$, and thus the conventional attention mechanism is not a proper way of calculating the causal effect.

To eliminate the spurious correlation brought by the hidden confounders, we should block the backdoor path between Z and Y: $Z \leftarrow X \leftrightarrow Y$. In this way, we can estimate the true causal effect between Z and Y, which is denoted as $P(Y|do(Z))$, where $do(\cdot)$ denotes the interventional operation [41]. We can cut off the link $X \rightarrow Z$ to block this backdoor path by stratifying the input variable X into different cases $\{\boldsymbol{x}\}$ and then measuring the average causal effects of Z on Y by the following expectation [43]:

$$P(Y|do(Z)) = \underbrace{\sum\nolimits_x P(X=x)}_{\text{CS-Sampling}} P(Y|X=x,Z), \tag{2}$$

where x denotes one possible input case. Here we denote it as Cross-Sample Sampling (**CS-Sampling**) since it comes from the other samples. Intuitively, CS-Sampling approximates the "physical intervention" which can break the spurious correlation caused by the hidden confounder. For example, the annotation "man-with-snowboard" is dominant in captioning dataset [19] and thus the predictor may learn the spurious correlation between the snowboard region with the word "man" without looking at the person region to reason what actually the gender is. CS-Sampling alleviates such spurious correlation by combining the person region with the other objects from other samples, *e.g*., bike, mirror, or brush, and inputting the combinations to the predictor. Then the predictor will not always see "man-with-snowboard" but see "man" with the other distinctive objects and thus it will be forced to infer the word "man" from the person region. With this deconfounded predictor, the selector will also be forced to select the legitimate evidence even we do not have any region-word supervisions.

By replacing $P(Y|\boldsymbol{z})$ in Eq. (1) by $P(Y|do(Z))$ in Eq. (2), we can calculate the true causal effect between X and Y:

$$\begin{aligned} &P(Y|do(X)) \\ &= \underbrace{\sum\nolimits_z P(Z=z|X)}_{\text{IS-Sampling}} \underbrace{\sum\nolimits_x P(X=x)}_{\text{CS-Sampling}} [P(Y|Z=z,X=x)]. \end{aligned} \tag{3}$$

This is also called the front-door adjustment, which is a fundamental causal inference technique for deconfounding the unobserved confounder [40]. Since our novel attention module is designed by using Eq. (3) as the training target, we name our attention module as *Causal Attention* (CATT).

3.2. In-Sample and Cross-Sample Attentions

To implement our causal attention (Eq. (3)) in a deep framework, we can parameterize the predictive distribution $P(Y|Z,X)$ as a network $g(\cdot)$ followed by a softmax layer since most vision-language tasks are transformed into classification formulations [63, 4]:

$$P(Y|Z,X) = \text{Softmax}[g(Z,X)]. \tag{4}$$

[1] For convenience, we simplify the notation of the backdoor path $X \leftarrow C \rightarrow M \rightarrow Y$ shown in Figure 3 to $X \leftrightarrow Y$.

As can be seen in Eq. (3), we need to sample X and Z, and feed them into the network to complete $P(Y|do(X))$. However, the cost of the network forward pass for all of these samples is prohibitively expensive. To address this challenge, we apply Normalized Weighted Geometric Mean (NWGM) approximation [66, 54] to absorb the outer sampling into the feature level and thus only need to forward the "absorbed input" in the network for once. Specifically, by NWGM approximation, IS-Sampling and CS-Sampling in Eq. (3) can be absorbed into the network as:

$$\begin{aligned} P(Y|do(X)) &\approx \text{Softmax}[g(\hat{\boldsymbol{Z}}, \hat{\boldsymbol{X}})], \\ \textbf{IS-Sampling:} \quad \hat{\boldsymbol{Z}} &= \sum\nolimits_z P(Z=z|h(X))\boldsymbol{z}, \\ \textbf{CS-Sampling:} \quad \hat{\boldsymbol{X}} &= \sum\nolimits_x P(X=x|f(X))\boldsymbol{x}. \end{aligned} \tag{5}$$

where $h(\cdot)$ and $f(\cdot)$ denote query embedding functions which can transform the input X into two query sets. Both of them can be parameterized as networks. Note that in a network, the variable X and Z are represented by embedding vectors, *e.g.*, an image region becomes an RoI representation, so we use bold symbols to signify these embedding vectors, *e.g.*, $\boldsymbol{z}$, $\boldsymbol{x}$ denote the embedding vectors of the variable z, x. $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{Z}}$ denote the estimations of the IS-Sampling and CS-Sampling, which can be packed into the matrix form [60]. The derivation details of Eq. (5) are given in the supplementary material.

Actually, the IS-Sampling estimation $\hat{\boldsymbol{Z}}$ is what a classic attention network calculates, which can be briefly expressed by the Q-K-V operation as the blue block in Figure 4:

$$\begin{aligned} \textbf{Input:} \quad & \boldsymbol{Q}_I, \boldsymbol{K}_I, \boldsymbol{V}_I, \\ \textbf{Prob:} \quad & \boldsymbol{A}_I = \text{Softmax}(\boldsymbol{Q}_I{}^T \boldsymbol{K}_I) \\ \textbf{Ouput:} \quad & \hat{\boldsymbol{Z}} = \boldsymbol{V}_I \boldsymbol{A}_I \end{aligned} \tag{6}$$

We denote Eq. (6) as the In-Sample attention (**IS-ATT**) and the subscript "I" emphasizes that it is estimating IS-Sampling. In this case, all the $\boldsymbol{K}_I$ and $\boldsymbol{V}_I$ come from the current input sample, *e.g.*, the RoI feature set. $\boldsymbol{Q}_I$ comes from $h(X)$, *e.g.*, in top-down attention, the query vector $\boldsymbol{q}_I$ is the embedding of the sentence context and in self-attention, the query set $\boldsymbol{Q}_I$ is also the RoI feature set. For $\boldsymbol{A}_I$, each attention vector $\boldsymbol{a}_I$ is the network estimation of the IS-Sampling probability $P(Z=z|h(X))$ and the output $\hat{\boldsymbol{Z}}$ is the estimated vector set of IS-Sampling in Eq. (5).

Inspired by Eq (6), we can also deploy a Q-K-V operation to estimate $\hat{\boldsymbol{X}}$ and name it as Cross-Sample attention (**CS-ATT**), which is the red block in Figure 4:

$$\begin{aligned} \textbf{Input:} \quad & \boldsymbol{Q}_C, \boldsymbol{K}_C, \boldsymbol{V}_C, \\ \textbf{Prob:} \quad & \boldsymbol{A}_C = \text{Softmax}(\boldsymbol{Q}_C{}^T \boldsymbol{K}_C), \\ \textbf{Ouput:} \quad & \hat{\boldsymbol{X}} = \boldsymbol{V}_C \boldsymbol{A}_C \end{aligned} \tag{7}$$

where $\boldsymbol{K}_C$, $\boldsymbol{V}_C$ come from the other samples in the training set, and $\boldsymbol{Q}_C$ comes from $f(X)$. In this case, $\boldsymbol{a}_C$ approximates $P(X=x|f(X))$ and $\hat{\boldsymbol{X}}$ is the CS-Sampling estimation in Eq. (5). In the implementations, we set $\boldsymbol{K}_C$ and $\boldsymbol{V}_C$ as the global dictionaries compressed from the whole training dataset since it is impossible to attend to all the samples in the training set. Specifically, we initialize this dictionary by using K-means over all the samples' embeddings in training set, *e.g.*, all the images' RoI features. In this way, $\boldsymbol{V}_C$ and $\boldsymbol{V}_I$ stay in the same representation space, which guarantees that the estimations of IS-Sampling and CS-Sampling: $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{X}}$ in Eq. (5) also have the same distribution.

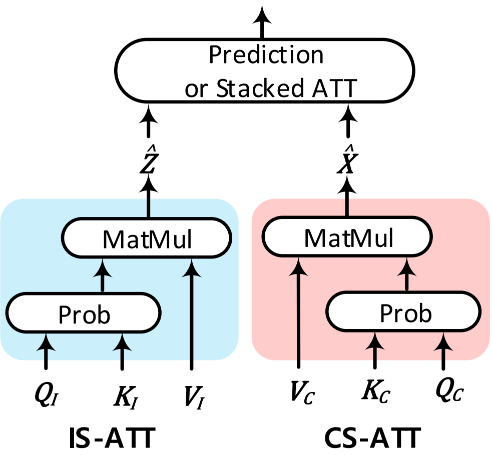

Figure 4. The sketch of a single causal attention module, which includes an IS-ATT (Eq. (6)) and a CS-ATT (Eq. (7)). After calculating $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{X}}$, we can input them into the predictor for making decisions or more stacked attention layers for further embedding.

To sum up, as shown in Figure 4, our single causal attention module estimates $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{X}}$ respectively by IS-ATT in Eq. (6) and CS-ATT in Eq. (7). After that, we can concatenate the outputs for estimating $P(Y|do(X))$ as in Eq. (5).

3.3. CATT in Stacked Attention Networks

In practice, attention modules can be stacked as deep networks, *e.g.*, the classic Transformer [60] or BERT architectures [16]. Our CATT can also be incorporated into these stacked attention networks and we experiment with Transformer [60] and LXMERT [57] in this paper. We briefly introduce their architectures here and discuss the implementation details in Section 4.2. Generally, our CATT replaces the first attention layer of these architectures to get the estimations of IS-Sampling $\hat{\boldsymbol{Z}}$ and CS-Sampling $\hat{\boldsymbol{X}}$, and then we input them into more attention layers for further embedding, as shown in Figure 4. For convenience, in these stacked attention networks, we still use IS-ATT and CS-ATT as the names of the attention modules to signify that this attention layer is dealing with the representations of the IS-Sampling or CS-Sampling.

Transformer+CATT. Figure 5 shows the architecture of

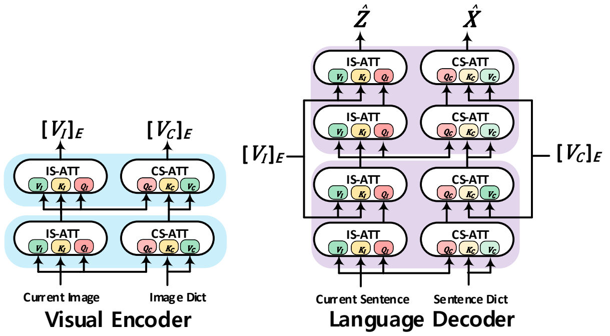

Figure 5. The Transformer+CATT architecture, which contains a visual encoder and a language decoder. We only show two layers in both parts for demonstrating how they are connected. In the implementations, both the encoder and decoder contain 6 layers. $[\boldsymbol{V}_I]_E$ and $[\boldsymbol{V}_C]_E$ denote the IS-ATT and CS-ATT outputs of the encoder, which are used as the inputs to the decoder. $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{X}}$ are the IS-ATT and CS-ATT outputs of the decoder, which are the estimations of IS-Sampling and CS-Sampling, respectively.

our vision-language Transformer+CATT. This architecture contains a vision encoder and a language decoder. In implementations, both the encoder and decoder contain 6 blue and purple blocks. The inputs of the encoder include the embedding set of the current image and a global image embedding dictionary. The IS-ATT and CS-ATT outputs of the encoder are input into the decoder for learning visiolinguistic representations. For the decoder, the inputs of the first IS-ATT and CS-ATT are respectively the current sentence embedding set and a global sentence embedding dictionary. The outputs of the decoder include two parts which respectively correspond to IS-Sampling $\hat{\boldsymbol{Z}}$ and CS-Sampling $\hat{\boldsymbol{X}}$, which will be concatenated and input into the final predictor. Importantly, by stacking many CATT layers, the estimated $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{X}}$ may not stay in the same representation space due to the non-convex operations in each attention module, *e.g.*, the position-wise feed-forward Networks [60]. To avoid this, we share the parameters of IS-ATT and CS-ATT in each CATT and then the outputs of them will always stay in the same representation space, where the detail formations are given in Eq. (8). As a result, the additional attention computation of CATT in LXMERT is $O(K*n)/O(n*n)$ at the first/other layer, where K is the size of the global dictionary and n is the number of word/image sequence.

LXMERT+CATT. Figure 6 demonstrates the architecture of our LXMERT+CATT, which contains three parts, a vision encoder with 5 self-CATT modules, a language encoder with 9 self-CATT modules, and a visiolinguistic decoder with 5 blocks where each one contains two cross-modality CATT (CM-CATT) and two self-CATT modules. For convenience, we merge the inputs (outputs) of IS-ATT and CS-ATT into one single line in (c). For example, the image inputs contain two parts which are the current image and a global image embedding dictionary. $[\boldsymbol{V}]_V$ ($[\boldsymbol{V}]_L$) denotes the visual (language) signal which also includes two parts $[\boldsymbol{V}_I]_V$ and $[\boldsymbol{V}_C]_V$ ($[\boldsymbol{V}_I]_L$ and $[\boldsymbol{V}_C]_L$) corresponding to IS-ATT and CS-ATT, respectively. Figure 6(b) sketches one cross-modality module used in the top part of the decoder in (c), where the visual signals are used as the queries in both IS-ATT and CS-ATT. Similar as the original LXMERT [57], we concatenate the outputs of both vision and language streams and input them into various predictors for solving different vision-language tasks. In implementations, we share the parameters of IS-ATT and CS-ATT in each causal attention module to force their outputs to have the same distributions.

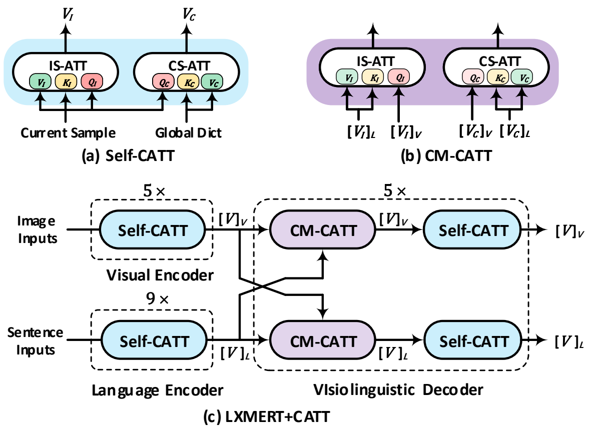

Figure 6. The architecture of LXMERT+CATT, which contains one visual encoder, one language encoder, and one visiolinguistic decoder. Note that each line in (c) contains two parts which respectively correspond to IS-ATT and CS-ATT. $[\boldsymbol{V}]_V$ and $[\boldsymbol{V}]_L$ denote the visual and language signals, respectively.

4. Experiments

We validated our Causal Attention (CATT) in three architectures for various vision-language tasks: Bottom-Up Top-Down (BUTD) LSTM [3] for Image Captioning (IC) [13, 35] and Visual Question Answering (VQA) [4], Transformer [60] for IC and VQA, and a large scale vision-language pre-training framework LXMERT [57] for VQA, Graph Question Answering (GQA) [22], and Natural Language for Visual Reasoning (NLVR) [55].

4.1. Datasets

MS COCO [13] has 123,287 images and each image is assigned with 5 captions. This dataset has two popular splits: the Karpathy split [23] and the official test split, which divide the whole dataset into $113,287/5,000/5,000$ and $82,783/40,504/40,775$ for training/validation/test, respectively. We used the Karpathy split to train the BUTD and Transformer based captioners and evaluate.

VQA2.0 [18] collects the images from MS COCO and assigns 3 questions for each image and 10 answers for each question. There are 80k/40k training/validation images available offline. We exploited the training set to train our BUTD and Transformer based VQA systems, and then evaluated the performances on three different splits: offline validation, online test-development, and online test-standard.

Pre-training and Fine-tuning Datasets for VLP. We followed LXMERT [57] to collect a large-scale vision-language pre-training dataset from the training and development sets of MS COCO, VQA2.0, GQA [22], and Visual Genome [26]. After collecting, this dataset contained 180K distinct images and 9.18M image-sentence pairs. We fine-tuned our VLP model on three tasks, which were VQA, GQA, and NLVR2 [55] and evaluated the performances on various test splits of them.

4.2. Implementation Details

The implementation details of BUTD+CATT and Transformer+CATT are given in C.4 of the supplementary material. Here we provide the details of LXMERT+CATT, which is the most significant experiments in this paper.

LXMERT + CATT. We used the architecture in Figure. 6 for large-scale vision-language pre-training. In this architecture, all IS-ATT and CS-ATT were deployed by 12-head scaled dot-product [60]:

$$\begin{aligned}
\textbf{Input:} \quad & \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \\
\textbf{Prob:} \quad & \boldsymbol{A}_i = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{W}_i^Q(\boldsymbol{K}\boldsymbol{W}_i^K)^T}{\sqrt{d}}) \\
\textbf{Single-Head}: \quad & \boldsymbol{H}_i = \boldsymbol{A}_i \boldsymbol{V} \boldsymbol{W}_i^V, \\
\textbf{Ouput:} \quad & \hat{\boldsymbol{V}} = \text{Embed}([\boldsymbol{H}_1, ..., \boldsymbol{H}_12]\boldsymbol{W}^H),
\end{aligned} \tag{8}$$

where $\boldsymbol{W}_i^*$ and $\boldsymbol{W}^H$ are all trainable matrices; $\boldsymbol{A}_i$ is the soft attention matrix for the i-th head; $[\cdot]$ denotes the concatenation operation, and Embed$(\cdot)$ means the feed-forward network and the residual operation as in [60]. The hidden size was set to 768. Importantly, we shared the parameters between IS-ATT and CS-ATT in each CATT to make the outputs stay in the same representation space. In this case, we also applied K-means to get the initializations and set the size of both dictionaries to 500. We extracted 36 RoI object features from each image by a Faster-RCNN [49] pre-trained on VG as in [3].

We followed the original LXMERT [57] to pre-train our LXMERT+CATT architecture by four tasks: masked cross-modality language modeling, masked object prediction, cross-modality image sentence matching, and image question answering. We used Adam optimizer with a linear-decayed learning rate schedule [16] where the peak learning rate was set to $5e^{-5}$. We pre-trained the model 20 epochs on 4 GTX 1080 Ti with a batch size of 192. The pre-training cost 10 days. To fairly compare the pre-training GPU hours with UNITER [14], we also carried an experiment by 4 V100 with batch size as 256 and it cost 6.5 days for pre-training. When fine-tuning the pre-trained model on VQA2.0, GQA, and NLVR2, the batch size was 32, training epochs was 4, and the learning rates were set to $5e^{-5}$, $1e^{-6}$, and $5e^{-5}$, respectively.

Table 1. The performances of various captioners on Karpathy split.

Models	B@4	M	R	C	S
BUTD [3]	37.2	27.5	57.3	125.3	21.1
LBPF [46]	38.3	**28.5**	58.4	127.6	22.0
GCN-LSTM [70]	38.2	**28.5**	58.3	127.6	22.0
SGAE [68]	38.4	28.4	**58.6**	127.8	**22.1**
BUTD+CATT	**38.6**	**28.5**	**58.6**	**128.3**	21.9
Transformer	38.6	28.5	58.4	128.5	22.0
VLP [78]	39.5	–	–	129.3	**23.2**
AoANet [21]	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$Transformer [15]	39.1	29.2	58.6	131.2	22.6
Transformer+CATT	39.4	**29.3**	**58.9**	**131.7**	22.8

Table 2. The bias degree of different models: "↑" and "↓" mean the higher the better and the lower the better, respectively. Red numbers denote the improvements after using our CATT modules.

Models	CHs↓	CHi↓	A@Gen↑	A@Attr↑	A@Act↑
BUTD	13.5	8.9	77%	41%	52%
BUTD+CATT	$10.7_{-2.8}$	$7.2_{-1.7}$	$85\%_{+8\%}$	$51\%_{+10\%}$	$60\%_{+8\%}$
Transformer	12.1	8.1	82%	47%	55%
Transformer+CATT	$9.7_{-2.4}$	$6.5_{-1.6}$	$92\%_{+10\%}$	$56\%_{+9\%}$	$64\%_{+9\%}$

4.3. Results and Analysis.

4.3.1 Image Captioning (IC)

Similarity Measurements. The results are reported in Table 1, where the top and bottom parts list various models which respectively deploy LSTM and Transformer as the backbones. In this table, B, M, R, C, and S denote BLEU [39], METEOR[6], ROUGE [31], CIDEr-D [61], and SPICE [2], respectively, which evaluate the similarities between the generated and the ground-truth captions.

Compared with two baselines BUTD and Transformer, we can find that BUTD+CATT and Transformer+CATT respectively achieve 3.0-point and 3.2-point improvements on CIDEr-D. More importantly, after incorporating our CATT modules into BUTD and Transformer, they have higher CIDEr-D scores than certain state-of-the-art captioners which deploy more complex techniques. For example, SGAE exploits scene graphs to transfer language inductive bias or $\mathcal{M}^2$Transformer learns multi-level visual relations though additional meshed memory networks. These comparisons suggest that our CATT module is a more powerful technique compared with the techniques used in these state-of-the-art captioners.

Bias Measurements. We measured the bias degree of the generated captions in Table 2 to validate that whether our CATT module can alleviate the dataset bias or not. In this table, CHs and CHi denote CHAIR$_s$ and CHAIR$_i$ [51], which are designed to measure the object bias. Apart from them, we also analyze three more specific biases: gender

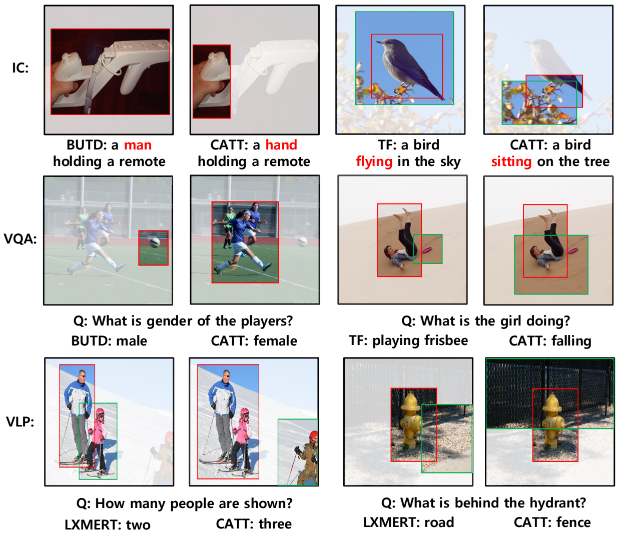
IC:
BUTD: a man holding a remote
CATT: a hand holding a remote
TF: a bird flying in the sky
CATT: a bird sitting on the tree
VQA:
Q: What is gender of the players?
BUTD: male
CATT: female
Q: What is the girl doing?
TF: playing frisbee
CATT: falling
VLP:
Q: How many people are shown?
LXMERT: two
CATT: three
Q: What is behind the hydrant?
LXMERT: road
CATT: fence

Figure 7. Six examples show that our CATT can correct different models to focus on the suitable regions. TF denotes Transformer. For BUTD, we show the region with the highest attention weight. For Transformer and VLP, the red region has the highest attention weight in top-down attention and the green region is the one most related to the red region in self-attention.

Table 3. Accuracy of various VQA models on different splits.

Models	loc-val	test-dev	test-std
BUTD [3]	63.20	65.32	65.67
MUTAN [8]	-	66.01	66.38
MLB [24]	65.07	66.27	66.62
BUTD+CATT	**65.12**	**67.13**	**67.26**
Transformer	66.29	69.53	69.82
DFAF [17]	66.21	70.22	70.34
MCAN [73]	67.20	70.63	70.90
TRRNet [67]	-	70.80	71.20
Transformer+CATT	**67.33**	**70.95**	**71.27**

Table 4. Accuracy of different question types on test-std split. Red numbers denote the improvements after using our CATT modules.

Models	*Yes/No*	*Number*	*Other*
BUTD	81.82	44.21	56.05
BUTD+CATT	$83.42_{+1.6}$	$48.96_{+4.75}$	$57.3_{+1.25}$
Transformer	86.25	50.7	59.9
Transformer+CATT	$87.40_{+1.15}$	$53.45_{+2.75}$	$61.3_{+1.4}$
LXMERT†	88.17	52.63	62.73
LXMERT+CATT	$88.6_{+0.43}$	$55.48_{+2.85}$	$63.39_{+0.66}$

bias, action bias, and attribute bias by calculating the accuracy of these words, which are denoted as A@Gen, A@Attr, and A@Act, respectively. From the results, we can see that after incorporating our CATT module, both BUTD and Transformer generate less biased captions, *e.g.*, the accuracies of gender, attribute, and action are respectively improved by 10%, 9%, and 9% when CATT is used in Transformer. The first row of Figure 7 shows two examples where BUTD and Transformer respectively attend to unsuitable regions and then generate incorrect captions, *e.g.*, BUTD attend to the remote region and infer the word "man" due to the dataset bias, while our CATT corrects this by attending to the hand region to generate the word "hand".

4.3.2 Visual Question Answering (VQA)

The top and bottom parts of Table 3 respectively report the performances of various LSTM and Transformer based VQA models, where loc-val, test-dev, and test-std denote the offline local validation, online test-development, and online test-standard splits. From this table, we can observe that after deploying our CATT module into BUTD and Transformer, the accuracies are consistently improved. More importantly, the deconfounded BUTD and Transformer outperform certain state-of-the-art models which are better than the original BUTD and Transformer.

Table 4 reports the accuracies of different question types on test-std split. It can be found that the accuracy of *number* has the largest improvements after using CATT modules, *i.e.*, 4.75-point and 2.75-point for BUTD and Transformer, respectively. Significantly, Transformer+CATT has a higher *number* accuracy than the large-scale pre-training model LXMERT: 53.45 vs. 52.63. As analyzed in [77], the counting ability depends heavily on the quality of the attention mechanism that a VQA model cannot correctly answer *number* questions without attending to all the queried objects. Thus the consistent improvements in *number* support that our CATT modules can largely ameliorate the quality of the conventional attention mechanism. The second row of Figure 7 shows that after incorporating CATT, BUTD and Transformer based VQA models can attend to the right regions for answering the questions.

4.3.3 Vision-Language Pre-training (VLP)

Table 5 shows the training burdens and the performances of various large-scale pre-training models on VQA2.0, GQA, and NLVR2. Note that LXMERT† and LXMERT respectively denote the results got from the officially released code and from the published paper. For ERNIE-VIL [72] and UNITER [14], they both have a BASE version and a LARGE version where BASE (LARGE) uses 12 (16) heads and 768 (1024) hidden units in multi-head product operations. We report the performances of their BASE versions since our model used 12 heads and 768 hidden units. For NLVR2, we report the performances of UNITER with the same Pair setting as our model.[2]

From this table, we can see that compared with LXMERT†, our LXMERT†+CATT respectively achieves 0.86, 1.23, 1.6-point improvements on the test-std splits of VQA2.0 and GQA and the test-P split of NLVR2. For example, compared with UNITER which uses fp16,

[2] The details of NLVR2 setting can be found in Table 5 of UNITER [14].

Table 5. Training burdens and performances of different large-scale vision-language pre-training models. “M” denotes million.

Models	Training Burdens		VQA2.0		GQA		NLVR2 (Pair)	
	GPU Hours	Image / Text	test-dev	test-std	test-dev	test-std	loc-val	test-P
LXMERT [57]	960 (titan xp)	0.18M / 9.18M	72.42	72.54	60.00	60.30	74.9	74.5
LXMERT† [57]	816 (1080Ti)	0.18M / 9.18M	71.96	72.18	59.90	59.94	74.8	74.4
ERNIE-VIL [72]	-	4.20M / 9.58M	72.62	72.85	-	-	-	-
UNITER [14]	882 (V100)	4.20M / 9.58M	72.80	72.91	-	-	75.85	75.80
12IN1 [33]	960 (V100)	5.40M / 7.48M	-	72.92	60.48	-	-	-
LXMERT†+CATT	960 (1080Ti), 624 (V100)	0.18M / 9.18M	72.81	73.04	60.84	61.17	76.40	76.00
LXMERT†+CATT↑	1536 (1080Ti), 1056 (V100)	0.18M / 9.18M	**73.54**	**73.63**	**61.87**	**62.07**	**77.27**	**77.23**

Table 6. The performances of various CATT ablation studies on the local validation sets. We show the CIDEr-D score for Image Captioning (IC) and the accuracies for the other tasks.

Models	Transformer		LXMERT		
	IC	VQA	VQA	GQA	NLVR2
Base	128.5	66.29	69.52	59.82	74.80
CATT w/o Init	129.8	66.56	69.81	60.14	75.22
CATT w/o Share	130.6	66.94	70.05	60.41	75.78
CATT+D#100	131.1	67.02	70.12	60.62	75.94
CATT+D#200	131.4	67.21	70.29	60.77	76.26
CATT+D#500	**131.7**	**67.33**	**70.40**	**60.90**	**76.40**

our LXMERT†+CATT uses fewer GPU hours and the pre-training data, while we have higher performances on VQA2.0: 73.04 vs. 72.91, and NLVR2: 76.0 vs. 75.80. Furthermore, inspired by UNITER [14], we enhanced our model and named this one as **LXMERT+CATT↑** by using conditional masking and more RoI features. Specifically, we extracted 64 RoI features from each image to guarantee that our model can be trained on 4 1080 Ti GPUs. It can be found that after using two insights from UNITER, our **LXMERT+CATT↑** can achieve higher performances than UNITER, even though we do not extract 100 RoI features for each image as them. These comparisons suggest that our CATT has great potential in large-scale VLP.

Also, as shown in Table 4, after incorporating CATT into LXMERT, we can observe that the accuracy of *Number* is further improved: 55.48 vs. 52.63, which suggests that our CATT improves the quality of the attention modules in VLP models. The third row of Figure 7 shows two examples where CATT modules correct LXMERT to focus on the right regions for answering the questions.

4.4. Ablation Studies

We carried exhaustive ablation studies to validate three variants of our causal attention module: K-means initialization, dictionary size, parameter sharing. In particular, we deployed these ablation studies in Transformer+CATT and LXMERT+CATT architectures.

Comparing Methods. **Base:** We denote the original Transformer and LXMERT architectures as Base. **CATT w/o Init:** CATT denotes the models introduced in Section 3.3. We did not use the K-means algorithm to initialize the global dictionaries but randomly initialized them. We shared the parameters between IS-ATT and CS-ATT. **CATT w/o Share:** We did not share the parameters between IS-ATT and CS-ATT. Here we used the K-means algorithm to initialize the dictionaries. **CATT+D#K:** We set the size of the global image and word embedding dictionaries to K by the K-means algorithm and shared the parameters between IS-ATT and CS-ATT.

Results and Analysis. Table 6 reports the performances of the ablation studies on the local validation sets. It can be found that after using our CATT, even without K-means initialization or parameter sharing, the performances are better than Base models. Also, we can observe that both K-means initialization and parameter sharing are useful for improving the performances, *e.g.*, CATT+D#500 outperforms both CATT w/o Init and CATT w/o Share. Such observation suggests that both strategies encourage the estimated IS-Sampling and CS-Sampling to stay in the same representation space, which is indeed beneficial in improving the performances. Also, by comparing the performances with different dictionary sizes, we can find that bigger dictionaries have better performances.

5. Conclusion

In this paper, we exploited the causal inference to analyze why the attention mechanism is easily misled by the dataset bias and then attend to unsuitable regions. We discovered that the attention mechanism is an improper approximation of the front-door adjustment and thus fails to capture the true causal effect between the input and target. Then a novel attention mechanism: causal attention (CATT) was proposed based on the front-door adjustment, which can improve the quality of the attention mechanism by alleviating the ever-elusive confounding effect. Specifically, CATT contains In-Sample and Cross-Sample attentions to estimate In-Sample and Cross-Sample samplings in the front-door adjustment and both of two attention networks abide by the Q-K-V operations. We implemented CATT into various popular attention-based vision-language models and the experimental results demonstrate that it can improve these models by considerable margins. In particular, CATT can promote a light VLP model comparable to a heavy one, which demonstrates its great potential in large-scale pre-training.

Acknowledgements. This work is partially supported by NTU Data Science and Artificial Intelligence Research Center (DSAIR), JRI, TIER2, and NTU-Alibaba Lab.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054, 2020. 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 6

[3] 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. 1, 2, 5, 6, 7

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3, 5

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1, 2

[6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[7] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 1

[8] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 7

[9] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019. 3

[10] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016. 3

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 1, 2

[12] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 1, 2020. 1, 2

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 5

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 1, 2, 6, 7, 8

[15] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 6

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2, 4, 6

[17] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019. 7

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 5

[19] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 2, 3

[20] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning, 2021. 3

[21] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *International Conference on Computer Vision*, 2019. 6

[22] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 3(8), 2019. 5, 6

[23] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 5

[24] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 7

[25] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017. 3

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome:

Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 3, 6

[27] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019. 3

[28] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. Sem{vlp}: Vision-language pre-training by aligning semantics at multiple levels, 2021. 2

[29] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[31] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 6

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 1, 2

[33] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 8

[34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016. 2

[35] Ruotian Luo. An image captioning codebase in pytorch, 2017. 5

[36] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2018. 1

[37] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020. 3

[38] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *arXiv preprint arXiv:2006.04315*, 2020. 3

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 6

[40] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. 2, 3

[41] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000. 2, 3

[42] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, pages 579–595, 2014. 2

[43] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 3

[44] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. 2, 3

[45] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10860–10869, 2020. 3

[46] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. Look back and predict forward in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8367–8375, 2019. 6

[47] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. 2

[48] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001. 2

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 6

[50] Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000. 1

[51] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 6

[52] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 2, 3

[53] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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 (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3

[54] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[55] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 5, 6

[56] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019. 2, 3

[57] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1, 2, 4, 5, 6, 8

[58] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 3716–3725, 2020. 3

[59] Roger BH Tootell, Nouchine Hadjikhani, E Kevin Hall, Sean Marrett, Wim Vanduffel, J Thomas Vaughan, and Anders M Dale. The retinotopy of visual spatial attention. *Neuron*, 21(6):1409–1422, 1998. 1

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2, 4, 5, 6

[61] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[62] Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020. 3

[63] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 3

[64] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 2

[65] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1, 2

[66] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 1, 2, 4

[67] Xiaofeng Yang, Guosheng Lin, Fengmao Lv, and Fayao Liu. Trrnet: Tiered relation reasoning for compositional visual question answering. 2020. 7

[68] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 6

[69] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 2

[70] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Computer Vision–ECCV 2018*, pages 711–727. Springer, 2018. 6

[71] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[72] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. 7, 8

[73] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 2, 7

[74] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 3

[75] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000*, 2020. 3

[76] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*, 2020. 2, 3

[77] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018. 7

[78] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019. 2, 6