6-2020

# Visual Commonsense R-CNN

Tan WANG

Jianqiang HUANG

Hanwang ZHANG

Qianru SUN
*Singapore Management University*, qianrusun@smu.edu.sg

# Visual Commonsense R-CNN

Tan Wang[1,3], Jianqiang Huang[2,3], Hanwang Zhang[3], Qianru Sun[4]
[1]University of Electronic Science and Technology of China [2]Damo Academy, Alibaba Group
[3]Nanyang Technological University [4]Singapore Management University
wangt97@hotmail.com, jianqiang.jqh@gmail.com, hanwangzhang@ntu.edu.sg, qianrusun@smu.edu.sg

## Abstract

*We present a novel unsupervised feature representation learning method, Visual Commonsense Region-based Convolutional Neural Network (**VC R-CNN**), to serve as an improved visual region encoder for high-level tasks such as captioning and VQA. Given a set of detected object regions in an image (e.g., using Faster R-CNN), like any other unsupervised feature learning methods (e.g., word2vec), the proxy training objective of VC R-CNN is to predict the contextual objects of a region. However, they are fundamentally different: the prediction of VC R-CNN is by using **causal intervention**: $P(Y|do(X))$, while others are by using the conventional **likelihood**: $P(Y|X)$. This is also the core reason why VC R-CNN can learn "sense-making" knowledge like* `chair` *can be sat — while not just "common" co-occurrences such as* `chair` *is likely to exist if* `table` *is observed. We extensively apply VC R-CNN features in prevailing models of three popular tasks: Image Captioning, VQA, and VCR, and observe consistent performance boosts across them, achieving many new state-of-the-arts*[1].

## 1. Introduction

> *"On the contrary, Watson, you can see everything. You fail, however, to reason from what you see."*
>
> *–Sherlock Holmes, The Adventure of the Blue Carbuncle*

Today's computer vision systems are good at telling us "what" (*e.g.*, classification [23, 31], segmentation [22, 39]) and "where" (*e.g.*, detection [54, 38], tracking [30, 34]), yet bad at knowing "why", *e.g.*, why is it `dog`? Note that the "why" here does not merely mean by asking for *visual* reasons — attributes like furry and four-legged — that are already well-addressed by machines; beyond, it also means by asking for high-level *commonsense* reasons — such as `dog` barks [17] — that are still elusive, even for us human philosophers [56, 21, 58], not to mention for machines.

[1] https://github.com/Wangt-CN/VC-R-CNN

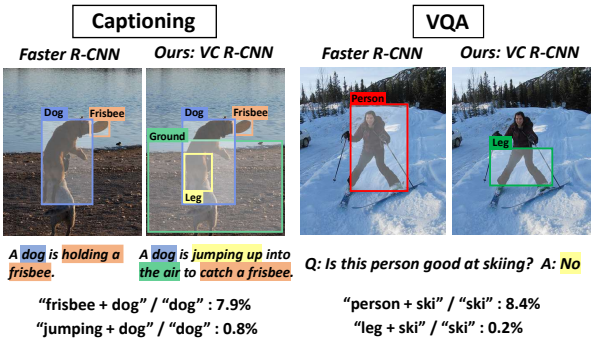



Figure 1. Examples of "cognitive errors" in image captioning and VQA due to the dataset bias. The ratio ./. denotes the co-occurrence% in ground-truth text (captioning: captions, VQA: questions). By comparing with the Faster R-CNN [54] based features [2], our VC R-CNN features can correct the errors, *e.g.*, more accurate visual relationships and visual attentions, by being more commonsense awareness.

It is not hard to spot the "cognitive errors" committed by machines due to the lack of common sense. As shown in Figure 1, by using only the visual features, *e.g.*, the prevailing Faster R-CNN [54] based Up-Down [2], machine usually fails to describe the exact visual relationships (the captioning example), or, even if the prediction is correct, the underlying visual attention is not reasonable (the VQA example). Previous works blame this for dataset bias without further justification [24, 44, 53, 7], *e.g.*, the large concept co-occurrence gap in Figure 1; but here we take a closer look at it by appreciating the difference between the "visual" and "commonsense" features. As the "visual" only tells "what"/"where" about `person` or `leg` *per se*, it is just a more descriptive symbol than its correspondent English word; when there is bias, *e.g.*, there are more `person` than `leg` regions co-occur with the word "ski", the visual attention is thus more likely to focus on the `person` region. On the other hand, if we could use the "commonsense" features, the action of "ski" can focuses on the `leg` region because of the common sense: we ski with legs.

We are certainly not the first to believe that visual features should include more commonsense knowledge, rather

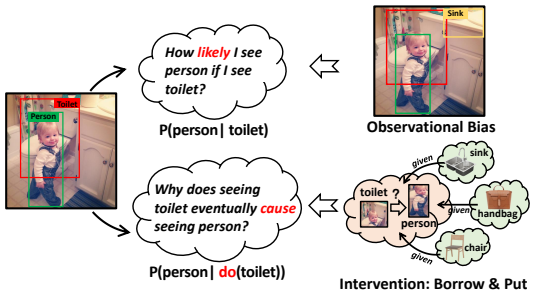



Figure 2. The illustration of why $P(Y|do(X))$ learns common sense while $P(Y|X)$ does not. Thanks to intervention, $P(Y|do(X))$ can "borrow" objects from other images and "put" them into the local image, to perform further justifications if $X$ truly causes $Y$ regardless of the unobserved confounders, and thus alleviate the observational bias.

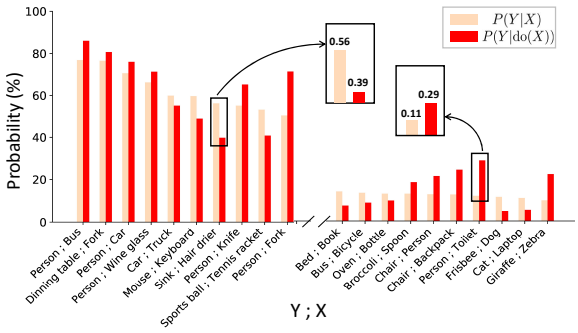



Figure 3. The sensible difference between the likelihood before (*i.e.*,$P(Y|X)$) and after intervention (*i.e.*, $P(Y|do(X))$) in MS-COCO. The object is represented by the 80 ground-truth class labels. Only 20 pairs are visualized to avoid clutter.

than just visual appearances. There is a trend in our community towards *weakly-supervised* learning features from large-scale vision-language corpus [41, 60, 61]. However, despite the major challenge in trading off between annotation cost and noisy multimodal pairs, common sense is not always recorded in text due to the reporting bias [66, 37], *e.g.*, most may say "people walking on road" but few will point out "people walking with legs". In fact, we humans naturally learn common sense in an *unsupervised fashion* by exploring the physical world, and we wish that machines can also imitate in this way.

A successful example is the unsupervised learning of word vectors in our sister NLP community [45, 12, 51]: a word representation $X$ is learned by predicting its contextual word $Y$, *i.e.*, $P(Y|X)$ in a neighborhood window. However, its counterpart in our own community, such as learning by predicting surrounding objects or parts [13, 43], is far from effective in down-stream tasks. The reason is that the commonsense knowledge, in the form of language sentences, has already been recorded in discourse; in contrast, once an image has been taken, the explicit knowledge why objects are contextualized will never be observed, so the true common sense that **causes** the existence of objects $X$ and $Y$ might be **confounded** by the spurious *observational bias*, *e.g.*, if `keyboard` and `mouse` are more often observed with `table` than any other objects, the underlying common sense that `keyboard` and `mouse` are parts of `computer` will be wrongly attributed to `table`.

Intrigued, we perform a toy MS-COCO [36] experiment with ground-truth object labels — by using a mental apparatus, *intervention*, that makes us human [50] — to screen out the existence of confounders and then eliminate their effect. We compare the difference between *association* $P(Y|X)$ and *causal intervention* $P(Y|do(X))$ [49]. Before we formally introduce *do* in Section 3.1, you can intuitively understand it as the following deliberate experiment illustrated in Figure 2: 1) "borrow" objects $Z$ from other images, 2) "put" them around $X$ and $Y$, then 3) test if $X$ still causes the existence of $Y$ given $Z$. The "borrow" and "put" is the spirit of intervention, implying that the chance of $Z$ is only dependent on us (probably subject to a prior), but independent on $X$ or $Y$. By doing so, as shown in Figure 3, $P(\texttt{sink}|do(\texttt{dryer}))$ is lower because the most common restroom context such as `towel` is forced to be seen as fair as others. Therefore, by using $P(Y|do(X))$ as the learning objective, the bias from the context will be alleviated.

More intrigued, $P(\texttt{person}|do(\texttt{toilet}))$ is higher. Indeed, `person` and `toilet` co-occur rarely due to privacy. However, human's *seeing* is fundamentally different from machine's because our instinct is to seek the *causality* behind any association [50] — and here comes the common sense. As opposed to the passive observation $P(Y|X)$: "How likely I see person if I see toilet", we keep asking "Why does seeing toilet eventually cause seeing person?" by using $P(Y|do(X))$. Thanks to intervention, we can increase $P(Y|do(X))$ by "borrowing" non-local context that might not be even in this image, for the example in Figure 2, objects usable by `person` such as `chair` and `handbag` — though less common in the restroom context — will be still fairly "borrowed" and "put" in the image together with the common `sink`. We will revisit this example formally in Section 3.1.

So far, we are ready to present our unsupervised region feature learning method: Visual Commonsense R-CNN (**VC R-CNN**), as illustrated in Figure 4, which uses Region-based Convolutional Neural Network (R-CNN) [54] as the visual backbone, and the causal intervention as the training objective. Besides its novel learning fashion, we also design a novel algorithm for the *do*-operation, which is an effective approximation for the imaginative intervention (cf. Section 3.2). The delivery of VC R-CNN is a region feature extractor for any region proposal, and thus it is fundamental and ready-to-use for many high-level vision tasks such as Image Captioning [68], VQA [3], and VCR [76]. Through extensive experiments in Section 5, VC R-CNN

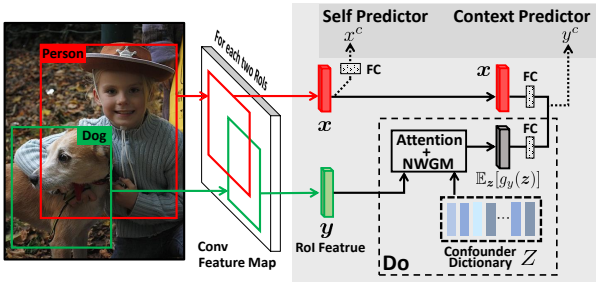



Figure 4. The overview of VC R-CNN. Any R-CNN backbone (*e.g.*, Faster R-CNN [54]) can be used to extract regions of interest (RoI) on the feature map. Each RoI is then fed into two sibling branches: a **Self Predictor** to predict its own class, *e.g.*, $x^c$, and a **Context Predictor** to predict its context labels, *e.g.*, $y^c$, with our **Do** calculus. The architecture is trained with a multi-task loss.

shows significant and consistent improvements over strong baselines — the prevailing methods in each task. Unlike the recent "Bert-like" methods [41, 60] that require huge GPU computing resource for pre-training features and fine-tuning tasks, VC R-CNN is light and non-intrusive. By "light", we mean that it is just as fast and memory-efficient as Faster R-CNN [54]; by "non-intrusive", we mean that re-writing the task network is not needed, all you need is `numpy.concatenate` and then ready to roll.

We apologize humbly to disclaim that VC R-CNN provides a philosophically correct definition of "visual common sense". We only attempt to step towards a **computational** definition in two intuitive folds: 1) common: unsupervised learning from the observed objects, and 2) sense-making: pursuing the causalities hidden in the observed objects. VC R-CNN not only re-thinks the conventional likelihood-based learning in our CV community, but also provides a promising direction — causal inference [50] — via practical experiments.

## 2. Related Work

**Multimodal Feature Learning.** With the recent success of pre-training language models (LM) [12, 10, 51] in NLP, several approaches [41, 60, 61, 9] seek weakly-supervised learning from large, unlabelled multi-modal data to encode visual-semantic knowledge. However, all these methods suffer from the reporting bias [66, 37] of language and the great memory cost for downstream fine-tuning. In contrast, our VC R-CNN is unsupervised learning only from images and the learned feature can be simply concatenated to the original representations.

**Un-/Self-supervised Visual Feature Learning** [14, 63, 43, 29, 77]. They aim to learn visual features through an elaborated proxy task such as denoising autoencoders [6, 67], context & rotation prediction [13, 18] and data augmentation [33]. The context prediction is learned from correlation while image rotation and augmentation can be regarded as applying the random controlled trial [50], which is active and non-observational (physical); by contrast, our VC R-CNN learns from the observational causal inference that is passive and observational (imaginative).

**Visual Common Sense.** Previous methods mainly fall into two folds: 1) learning from images with commonsense knowledge bases [66, 74, 57, 59, 69, 78] and 2) learning actions from videos [19]. However, the first one limits the common sense to the human-annotated knowledge, while the latter is essentially, again, learning from correlation.

**Causality in Vision.** There has been a growing amount of efforts in marrying complementary strengths of deep learning and causal reasoning [49, 48] and have been explored in several contexts, including image classification [8, 40], reinforcement learning [46, 11, 5] and adversarial learning [28, 26]. Lately, we are aware of some contemporary works on visual causality such as visual dialog [52], image captioning [73] and scene graph generation [62]. Different from their task-specific causal inference, VC R-CNN offers a generic feature extractor.

## 3. Sense-making by Intervention

We detail the core technical contribution in VC R-CNN: causal intervention and its implementation.

### 3.1. Causal Intervention

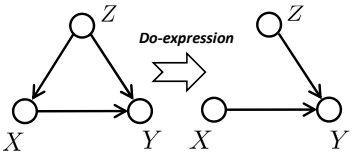



Figure 5. The causal intervention $P(Y|do(X))$. Nodes denote variables and arrows denote the direct causal effects.

As shown in Figure 5 (left), our visual world exists many confounders $z \in Z$ that affects (or causes) either $X$ or $Y$, leading to spurious correlations by only learning from the likelihood $P(Y|X)$. To see this, by using Bayes rule:

$$P(Y|X) = \sum\nolimits_z P\left(Y|X,z\right) \underline{P\left(z|X\right)}, \qquad (1)$$

where the confounder $Z$ introduces the observational bias via $P(z|X)$. For example, as recorded in Figure 6, when $P(z{=}\texttt{sink}|X{=}\texttt{toilet})$ is large while $P(z{=}\texttt{chair}|X{=}\texttt{toilet})$ is small, most of the likelihood sum in Eq. (1) will be credited to $P(Y{=}\texttt{person}|X{=}\texttt{toilet},z{=}\texttt{sink})$, other than $P(Y{=}\texttt{person}|X{=}\texttt{toilet},z{=}\texttt{chair})$, so, the prediction from `toilet` to `person` will be eventually focused on `sink` rather than `toilet` itself, *e.g.*, the learned features of a region `toilet` are merely its surrounding sink-like features.

As illustrated in Figure 5 (right), if we intervene $X$, *e.g.*, $do(X{=}\texttt{toilet})$, the causal link between $Z$ and $X$ is cut-off. By applying the Bayes rule on the new graph, we have:

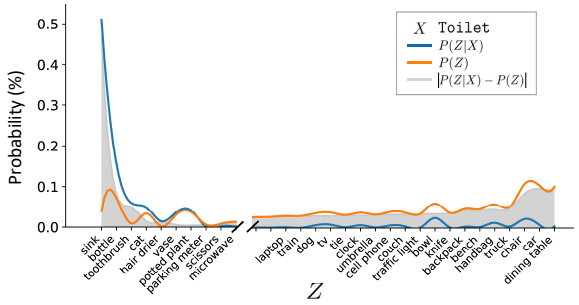


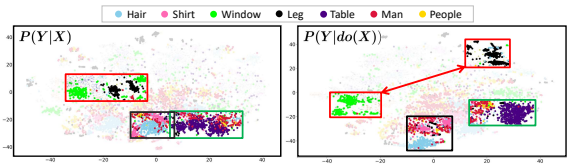

Figure 6. A case study of the differences between $P(z|\texttt{Toilet})$ and $P(z)$ from MS-COCO ground-truth object labels. Only 29 labels of $Z$ are shown to avoid clutter.

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

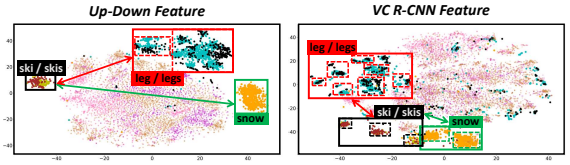
Compared to Eq. (1), $z$ is no longer affected by $X$, and thus the intervention deliberately forces $X$ to incorporate every $z$ fairly, subject to its prior $P(z)$, into the prediction of $Y$. Figure 6 shows the gap between the prior $P(z)$ and $P(z|\texttt{toilet})$, $z \in Z$ is the set of MS-COCO labels. We can use this figure to clearly explain the two interesting key results by performing intervention. Please note that $P(Y|X,z)$ remains the same in both Eq. (1) and Eq. (2),

Please recall Figure 3 for the sensible difference between $P(Y|X)$ and $P(Y|do(X))$. First, $P(\texttt{person}|do(\texttt{toilet}))>P(\texttt{person}|\texttt{toilet})$ is probably because the number of classes $z$ such that $P(z|\texttt{toilet})>P(z)$ is smaller than those such that of $P(z|\texttt{toilet}) < P(z)$, *i.e.*, the left grey area is smaller than the right grey area in Figure 6, making Eq. (1) smaller than Eq. (2). Second, we can see that $z$ making $P(z) < P(z|X)$ is mainly from the common restroom context such as `sink`, `bottle`, and `toothbrush`. Therefore, by using intervention $P(Y|do(X))$ as the feature learning objective, we can adjust between "common" and "sense-making", thus alleviate the observational bias.

Figure 7(a) visualizes the features extracted from MS-COCO images by using the proposed VC R-CNN. Promisingly, compared to $P(Y|X)$ (left), $P(Y|do(X))$ (right) successfully discovers some sensible common sense. For example, before intervention, `window` and `leg` features in red box are close due to the street view observational bias, *e.g.*, people walking on street with window buildings; after intervention, they are clearly separated. Interestingly, VC R-CNN `leg` features are closer to `head` while `window` features are closer to `wall`. Furthermore, Figure 7(b) shows the features of `ski`, `snow` and `leg` on same MS-COCO images via Up-Down (left) and our VC R-CNN (right). We can see the `ski` feature of our VC R-CNN is reasonably closer to `leg` and `snow` than Up-Down. Interestingly, VC R-CNN merges into sub-clusters (dashed boxes), implying that the common sense is actually multi-facet and varies from context to context.




(a) Object features learned by correlation $P(Y|X)$ and intervention $P(Y|do(X))$ (our VC R-CNN).




(b) Object features of Up-Down features and our VC R-CNN.

Figure 7. The t-SNE visualization [42] of object features trained on MS-COCO with Up-Down [2] provided Faster R-CNN labels. Features out of the label legend are faded out to avoid clutter.

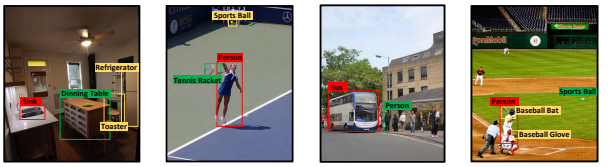
**$X \rightarrow Y$ or $Y \rightarrow X$?** We want to further clarify that both two causal directions between $X$ and $Y$ can be meaningful and indispensable with do calculus. For $X \rightarrow Y$, we want to learn the visual commonsense about $X$ (*e.g.*, `toilet`) that causes the existence of $Y$ (*e.g.*, `person`), and vice versa.

**Only objects are confounders?** No, some confounders are unobserved and beyond objects in visual commonsense learning, *e.g.*, color, attributes, and the nuanced scene contexts induced by them; however, in unsupervised learning, we can only exploit the objects. Fortunately, this is reasonable: 1) we can consider the objects as the partially observed children of the unobserved confounder [15]; 2) we propose the implementation below to approximate the contexts, *e.g.*, in Figure 8, `Stop sign` may be the child of the confounder "transportation", and `Toaster` and `Refrigerator` may contribute to "kitchen".

### 3.2. The Proposed Implementation

To implement the theoretical and imaginative intervention in Eq. (2), we propose the proxy task of predicting the local context labels of $Y$'s RoI. For the confounder set $Z$, since we can hardly collect all confounders in real world, we approximate it to a fixed confounder dictionary $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_N]$ in the shape of $N \times d$ matrix for practical use, where $N$ is the category size in dataset (*e.g.*, 80 in MS-COCO) and $d$ is the feature dimension of RoI. Each entry $\boldsymbol{z}_i$ is the averaged RoI feature of the $i$-th category samples in dataset. The feature is pre-trained by Faster R-CNN.

Specifically, given $X$'s RoI feature $\boldsymbol{x}$ and its contextual $Y$'s RoI whose class label is $y^c$, Eq. (2) can be implemented as $\sum_z P\left(y^c|\boldsymbol{x},\boldsymbol{z}\right)P\left(\boldsymbol{z}\right)$. The last layer of the network for label prediction is the Softmax layer: $P\left(y^c|\boldsymbol{x},\boldsymbol{z}\right) = \mathit{Softmax}(f_y(\boldsymbol{x},\boldsymbol{z}))$, where $f_y(\cdot)$ calculates the logits for $N$ categories, and the subscript $y$ denotes that $f(\cdot)$ is parameterized by $Y$'s RoI feature $\boldsymbol{y}$, motivated by the intuition that the prediction for $y^c$ should be characterized by $Y$. In




$X$: Sink $Y$: Dinning Table
$Z$: Toaster 0.75
Refrigerator 0.11
Microwave 0.01

$X$: Person $Y$: Tennis Racket
$Z$: Baseball Bat 0.78
Baseball Glove 0.05
Sports Ball 0.02

$X$: Bus $Y$: Person
$Z$: Stop Sign 0.97
Skateboard 0.002
Fire Hydrant 0.0005

$X$: Person $Y$: Sports Ball
$Z$: Baseball Glove 0.55
Sports Ball 0.32
Baseball Bat 0.09

Figure 8. The visualizations of the top 3 confounders given RoI feature $\boldsymbol{x}$ (red box) and $\boldsymbol{y}$ (green box), while numbers denote the attention weight. We can see that our model can recognize reasonable confounders $\boldsymbol{z}$, *e.g.*, the common context (yellow boxes).

summary, the implementation is defined as:

$$P\left(Y|do(X)\right) := \mathbb{E}_{\boldsymbol{z}}[Softmax(f_y(\boldsymbol{x},\boldsymbol{z}))]. \tag{3}$$

Note that $\mathbb{E}_{\boldsymbol{z}}$ requires expensive sampling.

**Normalized Weighted Geometric Mean (NWGM).** We apply NWGM [70] to approximate the above expectation. In a nutshell, NWGM[2] effeciently moves the outer expectation into the Softmax as:

$$\mathbb{E}_{\boldsymbol{z}}[Softmax(f_y(\boldsymbol{x},\boldsymbol{z}))] \overset{\text{NWGM}}{\approx} Softmax(\mathbb{E}_{\boldsymbol{z}}[f_y(\boldsymbol{x},\boldsymbol{z})]). \tag{4}$$

In this paper, we use the linear model $f_y(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{W}_2 \cdot g_y(\boldsymbol{z})$, where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{N\times d}$ denote the fully connected layer. Then the Eq. (4) can be derived as:

$$\mathbb{E}_{\boldsymbol{z}}[f_y(\boldsymbol{x},\boldsymbol{z})] = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{W}_2 \cdot \mathbb{E}_{\boldsymbol{z}}[g_y(\boldsymbol{z})]. \tag{5}$$

Note that the above approximation is reasonable, because the effect on $Y$ comes from both $X$ and confounder $Z$ (cf. the right Figure 5). Next, the key is to compute $\mathbb{E}_{\boldsymbol{z}}[g_y(\boldsymbol{z})]$.

**Computing $\mathbb{E}_{\boldsymbol{z}}[g_y(\boldsymbol{z})]$.** We encode $g_y(\cdot)$ as the Scaled Dot-Product Attention [64] to assign weights for different confounders in dictionary $\boldsymbol{Z}$ with specific $\boldsymbol{y}$. Specifically, given the $\boldsymbol{y}$ and confounder dictionary $\boldsymbol{Z}$, we can have $\mathbb{E}_{\boldsymbol{z}}[g_y(\boldsymbol{z})] = \sum_z [Softmax(\boldsymbol{q}^T\boldsymbol{K}/\sqrt{\sigma}) \odot \boldsymbol{Z}]P(\boldsymbol{z})$, where $\boldsymbol{q} = \boldsymbol{W}_3\boldsymbol{y}$, $\boldsymbol{K} = \boldsymbol{W}_4\boldsymbol{Z}^T$, $P(\boldsymbol{z})$ denotes the prior statistic probability and $\odot$ is the element-wise product, $\boldsymbol{W}_3$ and $\boldsymbol{W}_4$ are the embedding matrices that map each vector to the common subspace for similarity measure, $\sigma$ denotes the first dimension of $\boldsymbol{W}_3, \boldsymbol{W}_4$ as a constant scaling factor. Figure 8 visualizes the top 3 confounders ranked by the soft attention weights. Note that they are the cancer in learning "sense-making" features from $P(Y|X)$.

**Neural Causation Coefficient (NCC).** Due to the fact that the causality from the confounders as the category averaged features are not yet verified, that is, $\boldsymbol{Z}$ may contain colliders (or v-structure) [49] causing spurious correlations when intervention. To this end, we apply NCC [40] to remove possible colliders from $\boldsymbol{Z}$. Given $\boldsymbol{x}$ and $\boldsymbol{z}$, $NCC\left(\boldsymbol{x} \rightarrow \boldsymbol{z}\right)$ outputs the relative causality intensity from $\boldsymbol{x}$ to $\boldsymbol{z}$. Then we discard the training samples with strong collider causal intensities above a threshold.

[2]The detailed derivation about NWGM can be found in the Supp..

# 4. VC R-CNN

**Architecture.** Figure 4 illustrates the VC R-CNN architecture. VC R-CNN takes an image as input and generates feature map from a CNN backbone (*e.g.*, ResNet101 [23]). Then, unlike Faster R-CNN [54], we discard the Region Proposal Network (RPN). The ground-truth bounding boxes are directly utilized to extract the object level representation with the RoIAlign layer. Finally, each two RoI features $\boldsymbol{x}$ and $\boldsymbol{y}$ eventually branch into two sibling predictors: Self Predictor with a fully connected layer to estimate each object class, while Context Predictor with the approximated *do*-calculus in Eq. (3) to predict the context label.

**Training Objectives.** The Self-Predictor outputs a discrete probability distribution $p = (p[1], ..., p[N])$ over $N$ categories (note that we do not have the "background" class). The loss can be defined as $L_{self}(p, x^c) = -log(p[x^c])$, where $x^c$ is the ground-truth class of RoI $X$. The Context Predictor loss $L_{cxt}$ is defined for each two RoI feature vectors. Considering $X$ as the center object while $Y_i$ is one of the $K$ context objects with ground-truth label $y_i^c$, the loss is $L_{cxt}(p_i, y_i^c) = -log(p_i[y_i^c])$, where $p_i$ is calculated by $p_i = P(Y_i|do(X))$ in Eq. (3) and $p_i = (p_i[1], ..., p_i[N])$ is the probability over $N$ categories. Finally, the overall mulit-task loss for each RoI $X$ is:

$$L\left(X\right) = L_{self}(p, x^c) + \frac{1}{K}\sum\nolimits_i L_{cxt}(p_i, y_i^c). \tag{6}$$

**Feature Extractor.** We consider VC R-CNN as a visual commonsense feature extractor for any region proposal. Then the extracted features are directly concatenated to the original visual feature utilized in any downstream tasks. It is worth noting that we do NOT recommend early concatenations for some models that contain a self-attention architecture such as AoANet [25]. The reasons are two-fold. First, as the computation of these models are expensive, early concatenation significantly slows down the training. Second, which is more crucial, the self-attention essentially and implicitly applies $P(Y|X)$, which contradicts to causal intervention. We will detail this finding in Section 5.4.

# 5. Experiments

## 5.1. Datasets

We used the two following datasets for unsupervised learning VC R-CNN.

**MS-COCO Detection** [36]. It is a popular benchmark dataset for classification, detection and segmentation in our community. It contains 82,783, 40,504 and 40,775 images for training, validation and testing respectively with 80 annotated classes. Since there are 5K images from downstream image captioning task which can be also found in MS-COCO validation split, we removed those in training. Moreover, recall that our VC R-CNN relies on the context

| Model | Feature | MS-COCO | | | | Open Images | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | B4 | M | R | C | B4 | M | R | C |
| Up-Down | Origin [2] | 36.3 | 27.7 | 56.9 | 120.1 | 36.3 | 27.7 | 56.9 | 120.1 |
| | Obj | 36.7 | 27.8 | 57.5 | 122.3 | 36.7 | 27.8 | 57.5 | 122.3 |
| | Only VC | 34.5 | 27.1 | 56.5 | 115.2 | 35.1 | 27.2 | 56.6 | 115.7 |
| | +Det | 37.5 | 28.0 | 58.3 | 125.9 | 37.4 | 27.9 | 58.2 | 125.7 |
| | +Cor | 38.1 | 28.3 | 58.5 | 127.5 | 38.3 | 28.4 | 58.8 | 127.4 |
| | +VC | 39.5 | 29.0 | 59.0 | 130.5 | 39.1 | 28.8 | 59.0 | 130.0 |
| AoANet† | Origin[3] [25] | 38.9 | 28.9 | 58.8 | 128.4 | 38.9 | 28.9 | 58.8 | 128.4 |
| | Obj | 38.1 | 28.4 | 58.2 | 126.0 | 38.1 | 28.4 | 58.2 | 125.9 |
| | Only VC | 35.8 | 27.6 | 56.8 | 118.1 | 35.8 | 27.9 | 56.7 | 118.5 |
| | +Det | 38.8 | 28.8 | 58.7 | 128.0 | 38.7 | 28.6 | 58.7 | 127.7 |
| | +Cor | 38.8 | 28.9 | 58.7 | 128.6 | 38.9 | 28.8 | 58.7 | 128.2 |
| | +VC | **39.5** | **29.3** | **59.3** | **131.6** | 39.3 | 29.1 | 59.0 | 131.5 |
| SOTA | AoANet [25] | 38.9 | 29.2 | 58.2 | 129.8 | 38.9 | 29.2 | 58.2 | 129.8 |

Table 1. The image captioning performances of representative two models with ablative features on Karpathy split. The metrics: B4, M, R and C denote BLEU@4, METEOR, ROUGE-L and CIDEr-D respectively. The grey row highlight our features in each model. AoANet† indicates the AoANet without the refine encoder. Note that the Origin and Obj share the same results in MS-COCO and Open Images since they does not contain our new trained features.

prediction task, thus, we discarded images with only one annotated bounding box.

**Open Images [32].** We also used a much larger dataset called Open Images, a huge collection containing 16M bounding boxes across 1.9M images, making it the largest object detection dataset. We chose images with more than three annotations from the official training set, results in about 1.07 million images consisting of 500 classes.

## 5.2. Implementation Details

We trained our VC R-CNN on 4 Nvidia 1080Ti GPUs with a total batch size of 8 images for 220K iterations (each mini-batch has 2 images per GPU). The learning rate was set to 0.0005 which was decreased by 10 at 160K and 200K iteration. ResNet-101 was set to the image feature extraction backbone. We used SGD as the optimizer with weight decay of 0.0001 and momentum of 0.9 following [54]. To construct the confounder dictionary $\boldsymbol{Z}$, we first employed the pre-trained official ResNet-101 model on Faster R-CNN with ground-truth boxes as the input to extract the RoI features for each object. For training on Open Images, we first trained a vanilla Faster R-CNN model. Then $\boldsymbol{Z}$ is built by making average on RoIs of the same class and is fixed during the whole training stage.

## 5.3. Comparative Designs

To evaluate the effectiveness of our VC R-CNN feature (VC), we present three representative vision-and-language downstream tasks in our experiment. For each task, a **classic** model and a **state-of-the-art** model were both performed for comprehensive comparisons. For each method,

[3]Since we cannot achieve performances reported in original paper using the official code even with the help of author, here we show ours as the baseline. The original results can be found at the bottom row: SOTA.

| Model | BLEU-4 | | METEOR | | ROUGE-L | | CIDEr-D | |
|---|---|---|---|---|---|---|---|---|
| Metric | c5 | c40 | c5 | c40 | c5 | c40 | c5 | c40 |
| Up-Down [2] | 36.9 | 68.5 | 27.6 | 36.7 | 57.1 | 72.4 | 117.9 | 120.5 |
| SGAE [71] | 37.8 | 68.7 | 28.1 | 37 | 58.2 | 73.1 | 122.7 | 125.5 |
| CNM [72] | 37.9 | 68.4 | 28.1 | 36.9 | 58.3 | 72.9 | 123.0 | 125.3 |
| AoANet [25] | 37.3 | 68.1 | 28.3 | 37.2 | 57.9 | 72.8 | 124.0 | 126.2 |
| Up-Down+VC | 37.8 | 69.1 | 28.5 | 37.6 | 58.2 | 73.3 | 124.1 | 126.2 |
| AoANet†+VC | **38.4** | **69.9** | **28.8** | **38.0** | **58.6** | **73.8** | **125.5** | **128.1** |

Table 2. The performances of various single models on the online MS-COCO test server. Up-Down+VC and AoANet†+VC are the short for concatenated on [2] in Up-Down and AoANet†.

| Model | Feature | CHs | Chi | Model | Feature | CHs | Chi |
|---|---|---|---|---|---|---|---|
| Up-Down | Obj | 12.8 | 8.1 | AoANet† | Obj | 12.6 | 8.0 |
| | +Det | 12.0 | 7.5 | | +Det | 9.5 | 6.2 |
| | +Cor | 11.2 | 7.1 | | +Cor | 10.4 | 6.5 |
| | +VC | **10.3** | **6.5** | | +VC | **8.8** | **5.5** |

Table 3. Hallucination analysis [55] of various models on MS-COCO Karpathy test split to measure object hallucination for image captioning. The lower, the better.

we used the following five ablative feature settings: 1) **Obj**: the features based on Faster R-CNN, we adopted the popular used bottom-up feature [2]; 2) **Only VC**: pure VC features; 3) **+Det**: the features from training R-CNN with single self detection branch without Context Predictor. "+" denotes the extracted features are concatenated with the original feature, *e.g.*, bottom-up feature; 4) **+Cor**: the features from training R-CNN by predicting all context labels (*i.e.*, correlation) without the intervention; 5) **+VC**: our full feature with the proposed implemented intervention, concatenated to the original feature. For fair comparisons, we retained all the settings and random seeds in the downstream task models. Moreover, since some downstream models may have different settings in the original papers, we also quoted their results for clear comparison. For each downstream task, we detail the problem settings, dataset and evaluation metrics as below.

**Image Captioning.** Image captioning aims to generate textual description of an image. We trained and evaluated on the most popular "Karpathy" split built on MS-COCO dataset, where 5K images for validation, 5K for testing, and the rest for training. The sentences were tokenized and changed to lowercase. Words appearing less than 5 times were removed and each caption was trimmed to a maximum of 16 words. Five standard metrics were applied for evaluating the performances of the testing models: CIDEr-D [65], BLEU [47], METROT [4], ROUGE [35] and SPICE [1].

**Visual Question Answering (VQA).** The VQA task requires answering natural language questions according to the images. We evaluated the VQA model on VQA2.0 [20]. Compared with VQA1.0 [3], VQA2.0 has more question-image pairs for training (443,757) and validation (214,354), and all the question-answer pairs are balanced. Before training, we performed standard text pre-processing. Questions were trimed to a maximum of 14 words and candidate answer set was restricted to answers appearing more than 8

| Model | Feature | MS-COCO | | | | Open Images | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Y/N | Num | Other | All | Y/N | Num | Other | All |
| Up-Down | Obj [2] | 80.3 | 42.8 | 55.8 | 63.2 | 80.3 | 42.8 | 55.8 | 63.2 |
| | Only VC | 77.8 | 37.9 | 51.6 | 59.8 | 77.9 | 38.1 | 51.1 | 59.9 |
| | +Det | 81.8 | 44.5 | 56.8 | 64.5 | 81.9 | 44.7 | 56.5 | 64.6 |
| | +Cor | 81.5 | 44.6 | 57.1 | 64.7 | 81.3 | 44.7 | 57.0 | 64.6 |
| | +VC | 82.5 | 46.0 | 57.6 | 65.4 | 82.8 | 45.7 | 57.4 | 65.4 |
| MCAN | Obj [75] | 84.8 | **49.4** | 58.4 | 67.1 | 84.8 | **49.4** | 58.4 | 67.1 |
| | Only VC | 80.8 | 40.7 | 48.9 | 60.1 | 81.0 | 40.8 | 49.1 | 60.3 |
| | +Det | 84.8 | 49.2 | 58.8 | 67.2 | 84.9 | 49.3 | 58.4 | 67.2 |
| | +Cor | 85.0 | 49.2 | 58.9 | 67.4 | 85.1 | 49.1 | 58.6 | 67.3 |
| | +VC | **85.2** | **49.4** | **59.1** | **67.7** | 85.1 | 49.1 | 58.9 | 67.5 |
| SOTA | MCAN | 84.8 | **49.4** | 58.4 | 67.1 | 84.8 | **49.4** | 58.4 | 67.1 |

Table 4. Accuracy (%) of various ablative features on VQA2.0 validation set. Since the Obj achieves almost equal results with that in the original paper, here we just merge the two rows.

| Model | test-dev | | | | test-std |
|---|---|---|---|---|---|
| | Y/N | Num | Other | All | All |
| Up-Down [2] | 81.82 | 44.21 | 56.05 | 65.32 | 65.67 |
| BAN [27] | 85.46 | 50.66 | 60.50 | 69.66 | - |
| DFAF [16] | 86.09 | 53.32 | 60.49 | 70.22 | 70.34 |
| MCAN [75] | 86.82 | 54.04 | 60.52 | 70.63 | 70.90 |
| UP-Down+VC | 84.26 | 48.50 | 58.86 | 68.15 | 68.45 |
| MCAN+VC | **87.41** | **53.28** | **61.44** | **71.21** | **71.49** |

Table 5. Single model accuracies (%) on VQA2.0 test-dev and test set, where Up-Down+VC and MCAN+VC are the short for Object-VC R-CNN feature in Up-Down and MCAN.

times. The evaluation metrics consist of three pre-type accuracies (*i.e.*, "Yes/No", "Number" and "Other").

**Visual Commonsense Reasoning (VCR).** In VCR, given a challenging question about an image, machines need to present two sub-tasks: answer correctly (Q→A) and provide a rationale justifying its answer (QA→R). The VCR dataset [76] contains over 212K (training), 26K (validation) and 25K (testing) derived from 110K movie scenes. The model was evaluated in terms of 4-choice accuracy and the random guess accuracy on each sub-task is 25%.

### 5.4. Results and Analysis

**Results on Image Captioning.** We compared our VC representation with ablative features on two representative approaches: Up-Down [2] and AoANet [25]. For Up-Down model shown in Table 1, we can observe that with our +VC trained on MS-COCO, the model can even outperform current SOTA method AoANet over most of the metrics. However, only utilizing the pure VC feature (*i.e.*, Only VC) would hurt the model performance. The reason can be obvious. Even for human it is insufficient to merely know the common sense that "apple is edible" for specific tasks, we also need visual features containing objects and attributes (*e.g.*, "what color is the apple") which are encoded by previous representations. When comparing +VC with the +Det and +Cor without intervention, results also show absolute gains over all metrics, which demonstrates the effectiveness of our proposed causal intervention in representation learning. AoANet [25] proposed an "Attention on Attention" module on feature encoder and caption decoder for refining

| Model | Feature | MS-COCO | | Open Images | |
|---|---|---|---|---|---|
| | | Q→ A | QA→ R | Q→ A | QA→ R |
| R2C | Origin [76] | 63.8 | 67.2 | 63.8 | 67.2 |
| | Obj | 65.9 | 68.2 | 65.9 | 68.2 |
| | Only VC | 64.1 | 66.7 | 64.3 | 66.8 |
| | +Det | 66.1 | 68.5 | 66.1 | 68.3 |
| | +Cor | 66.5 | 68.9 | 66.6 | 69.1 |
| | +VC | 67.4 | 69.5 | 67.2 | 69.9 |
| ViLBERT† | Obj[3] | 69.1 | 69.6 | 69.1 | 69.6 |
| | Only VC | 68.8 | 70.1 | 68.9 | 70.1 |
| | +Det | 69.2 | 69.8 | 69.1 | 69.6 |
| | +Cor | 69.3 | 69.9 | 69.2 | 70.0 |
| | +VC | **69.5** | 70.2 | **69.5** | 70.3 |
| SOTA | ViLBERT† [41] | 69.3 | **71.0** | 69.3 | **71.0** |

Table 6. Experimental results on VCR with various visual features. ViLBERT† [41] denotes ViLBERT without pretraining process.

with the self-attention mechanism. In our experiment, we discarded the AoA refining encoder (*i.e.*, AoANet†) rather than using full AoANet since the self-attentive operation on feature can be viewed as an indiscriminate correlation against our do-expression. From Table 1 we can observe that our +VC with AoANet† achieves a new SOTA performance. We also evaluated our feature on the online COCO test server in Table 2. We can find our model also achieves the best single-model scores across all metrics outperforming previous methods significantly.

Moreover, since the existing metrics fall short to the dataset bias, we also applied a new metric CHAIR [55] to measure the object hallucination (*e.g.*, "hallucinate" objects not in image). The lower is better. As shown in Table 3, we can see that our VC feature performs the best on both standard and CHAIR metrics, thanks to our proposed intervention that can encode the visual commonsense knowledge.

**Results on VQA.** In Table 4, we applied our VC feature on classical Up-Down [2] and recent state-of-the-art method MCAN [75]. From the results, our proposed +VC outperforms all the other ablative representations on three answer types, achieving the state-of-the-art performance. However, compared to the image captioning, the gains on VQA with our VC feature are less significant. The potential reason lies in the limited ability of the current question understanding, which cannot be resolved by "visual" common sense. Table 5 reports the single model performance of various models on both test-dev and test-standard sets. Although our VC feature is limited by the question understanding, we still receive the absolute gains by just feature concatenation compared to previous methods with complicated module stack, which only achieves a slight improvement.

**Results on VCR.** We present two representative methods R2C [76] and ViLBERT [41] in this emerging task on the validation set. Note that as the R2C applies the ResNet backbone for residual feature extraction, here for fair comparison we switched it to the uniform bottom-up features. Moreover, for ViLBERT, since our VC features were not involved in the pretraining process on Conceptual Captions, here we utilized the ViLBERT† [41] rather than the full ViLBERT model. From the comparison with ablative visual


Figure 9. Qualitative examples of utilizing our VC feature (right) compared with using Obj feature (left). Boxes in images denote the attention region labeled with name and attention weight. Three rows represent Image Captioning, VQA and VCR task respectively.

| Component | Setting | CIDEr-D | Accuracy |
|---|---|---|---|
| Expectation | $\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}]$ | 128.9 | 67.2 |
| NCC | w/o NCC | 131.5 | **67.7** |
| Dictionary | Random Dictionary | 127.5 | 66.9 |
| | Context Dictionary | *Unstable Training* | |
| | Fixed Dictionary | **131.6** | **67.7** |

Table 7. Ablation studies of our proposed intervention trained on MS-COCO and evaluated with CIDEr-D (captioning) and Accuracy (VQA) on Karpathy testset and VQA2.0 validation set.

representations in Table 6, our +VC feature still shows the superior performances similar to the above two tasks.

**Results on Open Images.** To evaluate the transfer ability and flexibility of the learned visual commonsense feature, we also performed our proposed VC R-CNN on a large image detection collection. The results can be referred to Table 1&4&6. We can see that the performances are extremely close to the VC feature trained on MS-COCO, indicating the stability of our learned semantically meaningful representation. Moreover, while performing VCR with the dataset of movie clip, which has quite diverse distributions compared to the captioning and VQA built on MS-COCO, our VC R-CNN trained on Open Images achieves the reasonable better results.

### 5.5. Qualitative Analysis

We visualize several examples with our VC feature and previous Up-Down feature [2] for each task in Figure 9. Any other settings except for feature kept the same. We can observe that with our VC, models can choose more precise, reasonable attention area and explicable better performance.

### 5.6. Ablation Study

To evaluate our proposed intervention implementation, we carry out different settings for each module in our VC R-CNN and report results on captioning and VQA in Table 7. $\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}]$ denotes utilizing statistical $P(z)$ by counting from the dataset without attention. Random Dictionary denotes initializing the confounder dictionary by randomization rather than the average RoI feature, while the Context Dictionary encodes contexts in each image as a dynamic dictionary set. The default setting is the fixed confounder dictionary with our attention module and NCC, which gives the best results. We can observe that random dictionary and $\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}]$ would hurt the performance, which demonstrates the effectiveness of our implementation. Moreover, we can find that NCC refining just brings a little difference to the downstream task performance. The potential reason is that NCC just provides a qualitative prediction and may have deviation when applying on real-world visual feature. We will continue exploring NCC in the future work.

## 6. Conclusions

We presented a novel unsupervised feature representation learning method called VC R-CNN that can be based on any R-CNN framework, supporting a variety of high-level tasks by using only feature concatenation. The key novelty of VC R-CNN is that the learning objective is based on causal intervention, which is fundamentally different from the conventional likelihood. Extensive experiments on benchmarks showed impressive performance boosts on almost all the strong baselines and metrics. In future, we intend to study the potential of our VC R-CNN applied in other modalities such as video and 3D point cloud.

**Acknowledgments** We would like to thank all reviewers for their constructive comments. This work was partially supported by the NTU-Alibaba JRI and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

## References


[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer, 2016. 6

[2] 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, 4, 6, 7, 8

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2, 6

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACLW*, 2005. 6

[5] 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

[6] Yoshua Bengio, Eric Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *ICML*, 2014. 3

[7] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *NIPS*, 2019. 1

[8] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014. 3

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 3

[10] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*, July 2019. 3

[11] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, June 2019. 2, 3

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2, 3

[14] Justin Domke, Alap Karapurkar, and Yiannis Aloimonos. Who killed the directed model? In *CVPR*. IEEE, 2008. 3

[15] Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *AISTATS*, 2019. 4

[16] 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 *CVPR*, 2019. 7

[17] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2), 1977. 1

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3

[19] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 1, 2017. 3

[20] 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 *CVPR*, 2017. 6

[21] Ibrahim Abou Halloun and David Hestenes. Common sense concepts about motion. *American journal of physics*, 53(11), 1985. 1

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5

[24] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*. Springer, 2018. 1

[25] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019. 5, 6, 7

[26] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv preprint arXiv:1803.04929*, 2018. 3

[27] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. In *NIPS*, 2018. 7

[28] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *ICLR*, 2018. 3

[29] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, 2019. 3

[30] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *ICCVW*, 2015. 1

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 2020. 6

[33] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Rethinking data augmentation: Self-supervision and self-distillation. *arXiv preprint arXiv:1910.05872*, 2019. 3

[34] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 1

[35] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004. 6

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 2, 5

[37] Xiao Lin and Devi Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *CVPR*, 2015. 2, 3

[38] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*. Springer, 2016. 1

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[40] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, 2017. 3, 5

[41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*, 2019. 2, 3, 7

[42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov), 2008. 4

[43] Tomasz Malisiewicz and Alyosha Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009. 2, 3

[44] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *CVPR*, 2019. 1

[45] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 2

[46] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019. 3

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*. Association for Computational Linguistics, 2002. 6

[48] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4), 2014. 3

[49] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 3, 5

[50] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018. 2, 3

[51] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, June 2018. 2, 3

[52] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, 2020. 3

[53] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NIPS*, 2018. 1

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 3, 5, 6

[55] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 6, 7

[56] Sophia A Rosenfeld. *Common sense*. Harvard University Press, 2011. 1

[57] Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015. 3

[58] Barry Smith. The structures of the common-sense world. 1995. 1

[59] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. Learning visual knowledge memory networks for visual question answering. In *CVPR*, 2018. 3

[60] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2, 3

[61] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, Nov. 2019. 2, 3

[62] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 3

[63] Lucas Theis and Matthias Bethge. Generative image modeling using spatial lstms. In *NIPS*, 2015. 3

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 5

[65] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[66] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. Learning common sense through visual abstraction. In *ICCV*, 2015. 2, 3

[67] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*. ACM, 2008. 3

[68] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2

[69] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016. 3

[70] 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 *ICML*, 2015. 5

[71] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019. 6

[72] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *ICCV*, 2019. 6

[73] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 3

[74] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *NAACL*, 2016. 3

[75] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 7

[76] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 2, 7

[77] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, 2019. 3

[78] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*. Springer, 2014. 3