

Show, Deconfound and Tell: Image Captioning with Causal Inference

Bing Liu^{1*†}, Dong Wang^{1*}, Xu Yang², Yong Zhou^{1†}, Rui Yao¹, Zhiwen Shao¹, Jiaqi Zhao¹
¹School of Computer Science and Technology, China University of Mining and Technology,
²School of Computer Science and Engineering, Southeast University

{liubing, dongwang}@cumt.edu.cn, 101013120@seu.edu.cn,

{yzhou, ruiyao, zhiwen.shao, jiaqizhao}@cumt.edu.cn

Abstract

The transformer-based encoder-decoder framework has shown remarkable performance in image captioning. However, most transformer-based captioning methods ever overlook two kinds of elusive confounders: the visual confounder and the linguistic confounder, which generally lead to harmful bias, induce the spurious correlations during training, and degrade the model generalization. In this paper, we first use Structural Causal Models (SCMs) to show how two confounders damage the image captioning. Then we apply the backdoor adjustment to propose a novel causal inference based image captioning (CIIC) framework, which consists of an *interventional object detector (IOD)* and an *interventional transformer decoder (ITD)* to jointly confront both confounders. In the encoding stage, the *IOD* is able to disentangle the region-based visual features by deconfounding the visual confounder. In the decoding stage, the *ITD* introduces causal intervention into the transformer decoder and deconfounds the visual and linguistic confounders simultaneously. Two modules collaborate with each other to alleviate the spurious correlations caused by the unobserved confounders. When tested on MSCOCO, our proposal significantly outperforms the state-of-the-art encoder-decoder models on Karpathy split and online test split. Code is published in <https://github.com/CUMTGG/CIIC>.

1. Introduction

Image captioning aims to automatically understand the semantic information of an image and generate its accurate description. Inspired by neural machine translation [36], the encoder-decoder architecture has been widely adopted by most conventional image captioning models [2, 10, 39, 41], in which a deep convolutional neural network (CNN) serves

* Authors contributed equally.

† Corresponding author

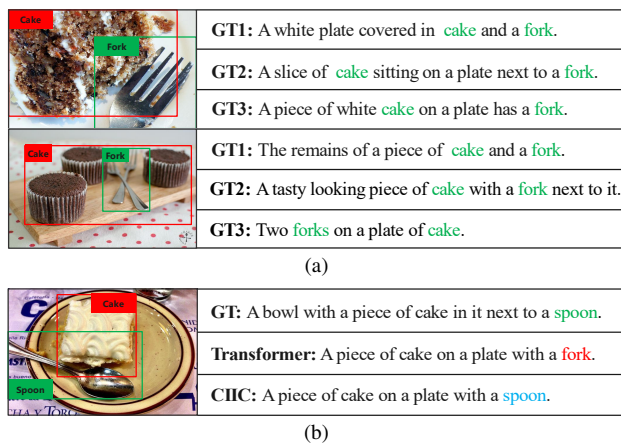


Figure 1. The example about the spurious correlation in image captioning. (a) Examples of the visual confounder (the visual feature of cake) and linguistic confounder (the word embedding of “cake”) in the MSCOCO training dataset when generating the word “fork”, where “GT1”, “GT2” and “GT3” denote three ground truth captions of each image chosen from the dataset. (b) Some captions generated by the original Transformer [37] and CIIC. The generated correct and incorrect words are colored by blue and red, respectively. “GT” means the ground truth caption.

as the encoder to extract visual features from the input image, and a recurrent neural network (RNN) is used as the decoder to generate the corresponding caption. Based on this architecture, a large number of improvements have been made by recent works, which mainly focus on two-fold: (i) Optimizing the visual representations of the input image [2, 15, 18, 42], and (ii) enhancing the architectural modeling capabilities for inter-modal and intra-modal interactions [8, 28].

In the aspect of visual representation, most captioning models apply a well-trained detector, *e.g.*, Faster R-CNN [33], to extract visual features. Nevertheless, these models neglected the problem of entangled visual features in the visual feature extraction stage. As shown in Figure 1a, the features of a region of the fork extracted by Faster R-

CNN tend to be its surrounding cake-like features since forks and cake co-occur too many times, *i.e.*, the feature representations of forks are severely affected by the visual feature of cake. In this case, the visual feature of the cake is actually one visual confounder, which builds a “short-cut path” [11] that leads to the spurious correlations between object features and target categories, *e.g.*, the learned cake-like features correspond to the class label of the fork. Consequently, it is critical to disentangle the visual features in the stage of visual representation to alleviate the spurious correlation between the cake region and the word “fork”.

In the aspect of model structure improvement, transformer-based models [13, 15, 17, 23, 27] have obtained the superior performance over the CNN-RNN based captioning methods. However, most transformer-based captioning models may still learn dataset bias caused by the hidden confounders. As shown in Figure 1a, when there are more forks than spoons co-occur with cake, due to both the visual confounder (*i.e.*, the visual feature of cake) and linguistic confounder (*i.e.*, the word embedding of “cake”), the traditional captioning models tend to learn the spurious correlation between the cake region and the word “fork” during training. Thus, as shown in Figure 1b, the original transformer usually generates the incorrect word “fork” instead of the correct word “spoon” for the test image.

Recently, Yang *et al.* [44] analyzed the spurious correlation between the visual features and captions by causal graph and proposed a deconfounded image captioning (DIC) framework to tackle the confounders. But they still have two limitations: (i) In their causal graph, the whole dataset is considered as the confounder and hard to be stratified. Thus, the complex front-door adjustment is utilized to deconfound it by introducing an additional mediator. (ii) DIC focuses on deconfounding the decoder while neglecting the confounded visual features in the encoder, leading to limited performance improvements.

To solve these problems, we first divide the confounder of the existing causal graph into two classes: visual confounder and linguistic confounder. Based on the detailed causal graph, we propose a novel causal inference based image captioning (CIIC) framework, which mainly consists of two components: an **interventional object detector (IOD)** and an **interventional Transformer decoder (ITD)** to jointly confront two kinds of confounders. Specifically, the **IOD** incorporates causal inference into Faster R-CNN [33] to cope with the visual confounder, aiming to obtain the disentangled region-based representations. The **ITD** implements causal intervention in the Transformer decoder by deconfounding both the visual and linguistic confounder simultaneously. As shown in Figure 1b, CIIC can effectively eliminate the spurious correlations caused by the visual and linguistic confounders and generate the correct word “spoon”.

Our contributions can be summarized as follows:

- We decompose the confounder into the visual and linguistic confounders and show a more detailed causal graph for the transformer-based image captioning system, which can be easily deconfounded by the backdoor adjustment, instead of the more complex front-door adjustment.
- We propose an **IOD** to disentangle the region-based features in the encoder and design a novel **ITD** by deconfounding the causal graph, which can effectively eliminate the spurious correlations caused by both visual and linguistic confounders.
- We implement our transformer-based **CIIC** framework to facilitate the unbiased captioning generation and extensively evaluate our approach on the MSCOCO benchmark [24]. CIIC achieves a new state-of-the-art performance compared to previous transformer-based captioning approaches.

2. Related Work

2.1. Image Captioning

The mainstream image captioning methods generally follow the encoder-decoder paradigm [2, 12, 39, 41, 50], where image features extracted by a CNN are fed into a recurrent net (often based on LSTM units) to generate the corresponding sentence. Since RNN-based models are limited by their sequential nature, the convolutional language model has been explored to replace conventional RNNs as well [3]. Different from the local operator essence of convolution, new transformer-based captioning models, based on the fully-attentive paradigm, have recently been proposed and achieved quite promising results [9, 13, 17, 25]. For example, spatial relationships between region features [13, 15] and relative geometry features between grids [27] were explicitly incorporated with geometric attention to enhance visual representations. Li *et al.* [23] introduced Entangled Attention that exploits visual and semantic information simultaneously. Pan *et al.* [28] applied Bilinear Pooling to encode region-level and image-level features. Despite great progress made on the basis of fully-attentive paradigms, how to cope with dataset biases caused by the visual and linguistic confounders in image captioning is still largely under-explored.

2.2. Causal Inference

Recently, some researchers have incorporated causal inference into deep learning models in the computer vision community [6, 26, 32, 47, 49, 51]. These efforts make it possible to endow DNNs with the abilities to learn the causal effects, which significantly advance the performances of many CV and NLP models, including image classification [4, 26], image semantic segmentation [47], visual feature representation [40], visual dialog [32], image captioning

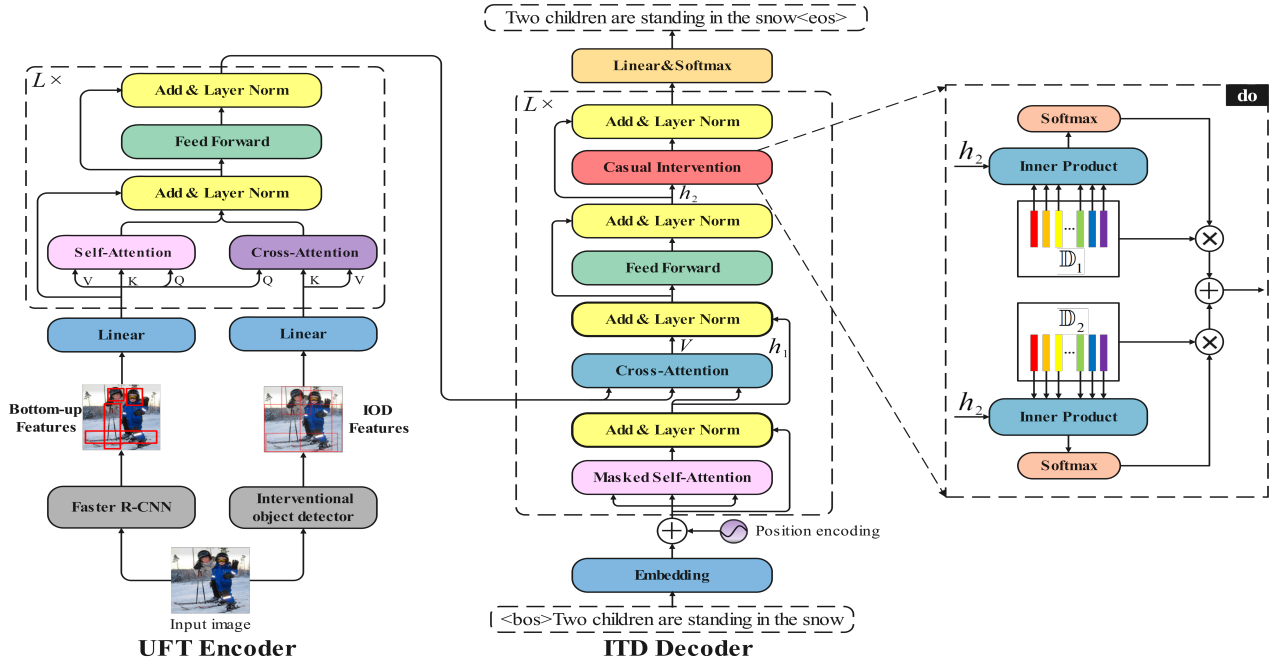


Figure 2. Illustration of the CIIC framework for image captioning. The RoI features are first disentangled by the interventional object detector, which are then combined with the bottom-up features of Faster R-CNN as inputs of the transformer encoder. In the decoder of CIIC, we propose a causal intervention module to confront both the visual and linguistic confounders for word prediction. The notation “ $L \times$ ” denotes that the block in the dotted box is stacked with L times. Our CIIC is able to effectively eliminate the spurious correlations that occurred in both the visual feature representation and caption generation to get more grounded image captions.

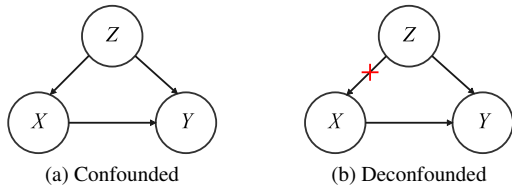


Figure 3. The causal intervention $P(Y|do(X))$ in object detection. The backdoor path $X \leftarrow Z \rightarrow Y$ is blocked by cutting off $Z \rightarrow X$, *i.e.*, the backdoor adjustment, which is able to effectively deconfound the unobserved confounder as a fundamental causal inference technique [30].

[44] and dialogue generation [51]. For example, Wang *et al.* [40] presented Visual Commonsense Region-based Convolutional Neural Network (VC R-CNN) to boost the performance of visual feature representation learning, in which causal intervention, instead of the conventional likelihood, is used to predict the contextual objects of a region. Yang *et al.* built a Deconfounded Image Captioning (DIC) framework [44] and a causal attention mechanism [45] respectively to cope with the confounders. However, these models still lack a detailed analysis of the confounders in both visual and linguistic domains. Consequently, we present the **IOD** to directly disentangle the visual features of the RoI proposals. The **IOD** incorporates causal intervention into self-predictor instead of the more complicated context predictor in VC R-CNN. As a result, the **IOD** can extract the

visual features from both single-object and multi-object images. Compared with the works by Yang *et al.* [44, 45], we devise the more detailed causal graphs for the transformer-based image captioning system and propose the **ITD** to simultaneously eliminate the spurious correlations caused by both the visual and linguistic confounders.

3. CIIC

As illustrated in Figure 2, CIIC is composed of a transformer encoder and a transformer decoder, in which causal inference is introduced into the visual representation step and sentence generation step, respectively.

3.1. Interventional Object Detector

Causal Intervention in Object Detection. In the causal graph [6, 26], a variable is defined as the confounder if such variable is a common cause for the other two variables. As illustrated in Figure 3a, we formulate the causalities among the region-based visual features X , the visual confounder Z of an image, and class labels Y based on SCM [6], where the direct edges represent the causalities between the two variables. On one hand, we denote the causal effect of Z on X as $Z \rightarrow X$ since the extracted visual features are inevitably affected by the visual contexts from the real world when the classifier of Faster R-CNN is trained. On the other

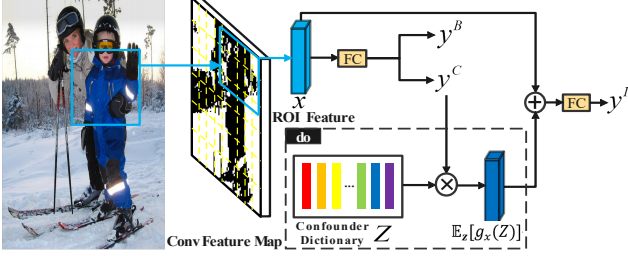


Figure 4. The architecture of our interventional object detector, where Faster R-CNN is used as the visual backbone [33] to extract visual features from regions of interest (RoIs). Subsequently, the extracted ROI feature is utilized to predict the class probability output y^C and the bounding box y^B , respectively. Depending on the class probability output y^C and the confounder dictionary Z , we perform causal intervention based on the **do** calculus to accurately predict the final object class label y^I .

hand, we have the causal effect $Z \rightarrow Y$ because the visual contexts also affect the classifier’s probability outputs. Hence, in the case of dataset bias, Faster R-CNN tends to learn some spurious correlation between X and Y caused by Z , *i.e.*, overexploiting the co-occurrence between the visual contexts and class labels to learn biased visual representations of image regions.

As shown in Figure 3a, conventional object detectors, such as Faster R-CNN, essentially use the likelihood $P(Y|X)$ as the training objective of classifier, which is usually affected by the confounder Z and gives rise to the spurious correlations. To see this, we formulate $P(Y|X)$ as:

$$P(Y|X) = \sum_z P(Y|X, Z = z) P(Z = z|X), \quad (1)$$

where the confounder Z generally brings about the observational bias via $P(z|X)$. For instance, when $P(z = \text{cake}|X = \text{fork})$ is large while $P(z = \text{spoon}|X = \text{fork})$ is small. According to Eq. (1), $P(Y = l_{\text{fork}}|X = \text{fork}, z = \text{cake})$, where l_{fork} denotes the class label of fork, plays a more important role than $P(Y = l_{\text{fork}}|X = \text{fork}, z = \text{spoon})$ in estimating $P(Y|X)$. Thus, the classifier learns the correlation between the visual features of cake and the class label of fork by mistake, *i.e.*, the learned ROI features of one fork are actually its surrounding cake-like visual features.

Motivated by the recent success of applying causal inference in deep learning [6, 26, 32, 47], we introduce causal intervention $P(Y|do(X))$ into object detection to block the backdoor path $X \leftarrow Z \rightarrow Y$, where the do calculus **do**(•) plays a role of cutting off $Z \rightarrow X$. As shown in Figure 3b, the backdoor adjustment is exploited to achieve $P(Y|do(X))$ as follows [40]:

$$P(Y|do(X)) = \sum_z P(Y|X, Z = z) P(Z = z). \quad (2)$$

In Eq. (2), $P(Y|do(X))$ forces X to fairly “borrow” each z in the confounder set and “put” them together for the

prediction of Y . In this way, the classifier removes the confounding effect and learns the true causality from X to Y , leading to the visual representations of high quality. However, Eq. (2) requires expensive sampling to estimate $P(Y|do(X))$ when applying it to a deep object detection network, which will make training time prohibitive. Fortunately, by applying the Normalized Weighted Geometric Mean (NWGM) approximation [40, 41], Eq. (2) can be approximated as:

$$P(Y|do(X = x)) \approx P\left(Y|\text{concat}\left(x, \frac{1}{n} \sum_{i=1}^n P(y_i^c|x) z_i\right)\right), \quad (3)$$

where $\text{concat}(\cdot)$ denotes vector concatenation, y_i^c is the i -th class label and $P(y_i^c|x)$ is the pre-trained classifier’s probability output that x belongs to class y_i^c . Note that we approximate the confounder in Eq. (2) to a predefined confounder dictionary $Z = [z_1, z_2, \dots, z_n]$, where n is the class number and $z_i \in \mathbb{R}^d$ denotes the average ROI feature of the i -th class pretrained by Faster R-CNN.

IOD Architecture. In Figure 4, we propose a novel **IOD** network to extract the disentangled visual features, where Faster R-CNN [33] is used as the visual backbone. In IOD, we use the same bounding box regressor as Faster R-CNN to specify each ROI on the feature map. As shown in Figure 4, the ROI feature x is then fed into two parallel branches to predict the class probability output y^C and bounding box y^B , respectively. Finally, based on the ROI feature x , the class probability output y^C , and the predefined confounder dictionary Z , we make the do calculus to implement the interventional class predictor and output the final object class label y^I , namely, the **IOD** applies Eq. (3) as the new classification objective to replace the classifier of Faster R-CNN. In this way, the ROI feature x can be effectively disentangled and subsequently adopted to facilitate the Transformer decoder to generate the unbiased caption.

3.2. Transformer Encoder with Multi-view Visual Representation

Now we are ready to utilize the IOD to extract the disentangled object features (called the IOD features) from any ROI proposal. Considering the bottom-up visual features obtained by the Up-Down approach [2] have the discriminative ability of different object attributes, we integrate the IOD features with the bottom-up features extracted from the same image to facilitate the visual representation of the CIIC model. Since the bottom-up and IOD features are unaligned, we introduce a multi-view transformer encoder, namely, Unaligned Feature Transformer (UFT) encoder, to adapt them.

As illustrated in Figure 2, the UFT encoder takes the unaligned visual features as inputs and performs the alignment and fusion operations simultaneously. Assume that

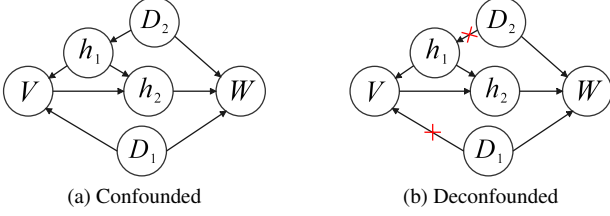


Figure 5. The causal intervention $P(W|do(V), do(h_1))$ in image captioning. Aiming at capturing the true causal effect: $V \rightarrow W$, we block the backdoor path $V \leftarrow h_1 \leftarrow D_2 \rightarrow W$ and $V \leftarrow D_1 \rightarrow W$ by cutting off $D_2 \rightarrow h_1$ and $D_1 \rightarrow V$ simultaneously.

the extracted bottom-up features and IOD features from an image can be respectively denoted as $X_F \in \mathbb{R}^{m \times d_1}$ and $X_I \in \mathbb{R}^{n \times d_2}$, where $m \neq n$ and $d_1 \neq d_2$. Two linear layers are utilized to transform X_F and X_I into a common d -dimensional space, denoted as \tilde{X}_F and \tilde{X}_I , respectively. Subsequently, we choose \tilde{X}_F as the features of the primary view and exploit it to learn the cross-attentions over \tilde{X}_I :

$$\tilde{X}_I = \text{MultiHead}(\tilde{X}_F, \tilde{X}_I, \tilde{X}_I), \quad (4)$$

where $\text{MultiHead}(\cdot)$ denotes the multi-head attention function of standard transformer, and $\tilde{X}_I \in \mathbb{R}^{m \times d}$ is the attended features over \tilde{X}_I . Accordingly, we model the multi-head self-attentions within \tilde{X}_F as follows:

$$\tilde{X}_F = \text{MultiHead}(\tilde{X}_F, \tilde{X}_F, \tilde{X}_F). \quad (5)$$

Note that \tilde{X}_I has the same shape as \tilde{X}_F and \tilde{X}_F , we encapsulate them by the AddNorm operator as follows:

$$\mathcal{F} = \text{LayerNorm}(\tilde{X}_F + \tilde{X}_F + \tilde{X}_I), \quad (6)$$

where $\text{LayerNorm}(\cdot)$ denotes layer normalization [37]. Finally, the fused features \mathcal{F} are fed into the FFN module to generate the encoding results of UFT. Noticeably, the UFT encoder is actually stacked in depth to generate more abstract and discriminative visual features for decoding, we omit them for a concise expression.

3.3. Interventional Transformer Decoder

To alleviate the spurious correlations between attended visual features and their corresponding words, we build a novel Transformer-based decoder architecture, which incorporates a causal intervention module into each Transformer decoder layer to cope with both the visual and linguistic confounders in image captioning.

Causal Intervention in Image Captioning. We first formulate the causalities among the attended visual features V , visual context D_1 , linguistic context D_2 , attended word features over the partially generated sentence h_1 , fused feature h_2 and predicted word W with an SCM, as illustrated in Figure 5a. Concretely, the causal effect $V \rightarrow W$ denotes that the attended visual features cause the generation of their corresponding words. The causal effect of D_1 on V stands for $D_1 \rightarrow V$ because the attended visual features are severely affected by some frequently appearing visual con-

texts when one captioner is trained, while the causal effect $D_1 \rightarrow W$ means that the visual contexts directly affect the frequency of some related words in the captions. In addition, $D_2 \rightarrow h_1 \rightarrow V$ denotes that the attended word features, affected by the linguistic contexts, are used to guide the attended visual features via multi-head cross-attention. $h_1 \rightarrow h_2$, $V \rightarrow h_2$ and $h_2 \rightarrow W$ indicate that the decoder integrates the visual features with linguistic features and utilizes the fused feature h_2 to infer the next word W . Thus, when we use the observational likelihood $P(W|V, h_1)$ as the training target, the captioner is likely to learn some spurious correlation between V and W due to the confounders D_1 and D_2 . To describe the principle of causal intervention in image captioning, we formulate $P(W|V, h_1)$ as:

$$P(W|V, h_1) = \sum_{d_2} P(d_2|h_1) \cdot \sum_{d_1} P(W|V, h_1, d_1, d_2)P(d_1|V), \quad (7)$$

where the confounders D_1 and D_2 generally introduces the observational biases via $P(d_1|V)$ and $P(d_2|h_1)$. Similar to the IOD, we substitute causal intervention $P(W|do(V), do(h_1))$ for the conventional training objective of image captioning, aiming at removing the causal effects of D_1 on V and D_2 on h_1 , as shown in Figure 5b. Thus, the two backdoor paths: $V \leftarrow D_1 \rightarrow W$ and $h_1 \leftarrow D_2 \rightarrow W$ are blocked and the spurious correlations are eliminated. Suppose that the confounders D_1 and D_2 can be stratified respectively, $P(W|do(V), do(h_1))$ can be calculated based on the backdoor adjustment as follows [6, 26]:

$$P(W|do(V), do(h_1)) = \sum_{d_2} P(d_2) \sum_{d_1} P(W|V, h_1, d_1, d_2) P(d_1). \quad (8)$$

Consequently, based on the interventional probability in Eq. (8), the image captioner is forced to learn the true causal effect: $V \rightarrow W$ rather than the spurious correlations caused by the visual confounder D_1 and linguistic confounder D_2 .

Likewise, we build the approximate visual confounder dictionary \mathbb{D}_1 and linguistic confounder dictionary \mathbb{D}_2 since both D_1 and D_2 are unobserved and beyond objects in image captioning. On the one hand, we construct the visual matrix $V_r \in \mathbb{R}^{c \times d_v}$ by setting each entry as the average RoI feature of objects in each class, where c denotes the class size and d_v is the dimensionality of each RoI feature. On the other hand, a set of d_e -dimensional word embeddings $W_e \in \mathbb{R}^{N \times d_e}$ from a pre-defined word vocabulary are utilized to build a semantic space. Then, the captioner is trained to learn two linear projections $P_v \in \mathbb{R}^{d_v \times d}$ and $P_w \in \mathbb{R}^{d_e \times d}$ to respectively transform V_r and W_e into \mathbb{D}_1 and \mathbb{D}_2 , i.e., $\mathbb{D}_1 = V_r P_v$, $\mathbb{D}_2 = W_e P_w$. Thus, Eq. (8) can be computed by utilizing the NWGM approximation [40, 41] as follows:

$$P(W|do(V), do(h_1)) \approx \text{Softmax}\{g(h_2, \mathbb{E}_{D_1}[D_1], \mathbb{E}_{D_2}[D_2])\}, \quad (9)$$

where $g(\cdot)$ represents an FC layer, $\mathbb{E}_{D_1}[D_1] \approx \text{softmax}(\mathbb{D}_1 h_2) \mathbb{D}_1$ and $\mathbb{E}_{D_2}[D_2] \approx \text{softmax}(\mathbb{D}_2 h_2) \mathbb{D}_2$. Similar to DIC [44], we set D_1 and D_2 to be conditioned on the fused feature h_2 to increase the representation power of ITD.

Transformer Decoder Architecture. The flowchart of the Transformer decoder architecture is shown in Figure 2. Similar to the Transformer encoder, the decoder consists of L identical decoder layers stacked in sequence. Unlike the original Transformer decoder, our proposal inserts a CI module after the FFN module. Specifically, by means of the visual dictionary \mathbb{D}_1 and linguistic dictionary \mathbb{D}_2 , the CI module integrates the fused feature h_2 with the expectations of both the visual confounder D_1 and linguistic confounder D_2 to predict the next word, *i.e.*, it actually implements causal intervention by the backdoor adjustment according to Eq. (9). The output of the last decoder layer is subsequently projected into an N -dimensional space by a linear embedding layer, where N is the vocabulary size. Finally, a softmax operation is adopted to predict a probability over words in the vocabulary.

3.4. Training Details

Following the same training strategies in [2, 10], our model is first pre-trained with the word-level cross entropy (XE) loss:

$$L_{XE}(\theta) = - \sum_{t=1}^T \log(p_{\theta}(w_t^* | do(V), do(h_1), w_{1:t-1}^*)), \quad (10)$$

where θ denotes the parameters of CIIC, $w_{1:T}^*$ is the target ground truth sequence.

After that, the model is optimized for the non-differentiable metric via Reinforcement Learning (RL). In practice, we adopt a variant of the Self-Critic Sequence Training (SCST) [34] on sequences sampled by beam search. The target is to minimize the following negative expected score:

$$L_{RL}(\theta) = -\mathbb{E}_{w_{1:T} \sim p_{\theta}}[r(w_{1:T})], \quad (11)$$

where the reward $r(\cdot)$ denotes the CIDEr-D score.

In the testing stage, we use beam search to generate the sentence word-by-word and obtain the sequence with the largest probability among those in the last beam.

4. Experiments

4.1. Experimental Setup

MS COCO Dataset [24]. This popular benchmark dataset contains 123,287 images and each of them equipped with 5 manually annotated sentences. In the experiments, We adopt two popular splits: the Karpathy split [20] and the online test split.

Evaluation Metrics. To evaluate the performance of different captioning methods, we utilize the full set of the stan-

dard evaluation metrics, including BLEU [29], METEOR [5], ROUGR [7], CIDEr [38], and SPICE [1]. Besides, we employ two metrics: CHAIR_s and CHAIR_i [35] to measure the object bias degree of the generated captions.

Implementation Details. To represent the image features, we first train the proposed IOD on the MSCOCO dataset to extract the 1024 dimensional IOD features of the top-100 objects with the highest confidence scores. Then, we use the pre-trained Up-Down model [2] to extract the 2048 dimensional bottom-up features of the detected objects. Finally, we linearly project both features to the 512 dimensional vectors and fed them into the UFT encoder. To represent the words, we respectively utilize one-hot vectors and pre-trained GloVe word embeddings [31] in the experiments. Both of them are linearly projected to the 512 dimensional input vectors of ITD. To represent word positions inside the sentence, we sum the input vectors and their sinusoidal positional encodings [36] before the first decoding layer.

Following the same settings in [37], we convert all sentences to lowercase, delete the punctuation characters and tokenize each caption. We construct a new vocabulary by selecting the words which appear more than 5 times. In addition, we use 8 attentive heads in both the encoder and decoder of CIIC. The latent dimensionality in each head is set to $d_h = d/h = 64$, where the latent dimensionality d is 512. We train our CIIC on one Nvidia 3080 GPU with a batch size of 10 images for 220K iterations. For fair comparisons, we employ ResNet-101 as the backbone for both the image feature extraction and encoding.

In the training stage, we employ the Adam optimizer [21] with a batch size of 10 and 20000 warmup steps. Our models are first trained for 30 epochs based on the XE loss and then optimized with the SCST approach [34] for additional 30 epochs with a beam size of 5. The learning rate is set to 5×10^{-6} . During the inference stage, the beam search [36] is also adopted with a beam size of 3.

4.2. Ablation Studies

We carry out extensive experiments to investigate the impacts of different modules on captioning performance.

Comparing Methods. Base: We denote the original Transformer captioning model with the bottom-up features as Base. **Base+GloVe:** We incorporate the GloVe embedding [31] into Base to represent the words. **Base+ITD:** We introduce the ITD module into the decoder of Base. **Base+ITD+GloVe:** Compared with Base+ITD, we further integrate the GloVe embedding with Base. **Base+UFT:** We exploit the UFT to improve the visual representation ability of Base. **Base+UFT+GloVe:** Compared with Base+UFT, we replace the one-hot embeddings by the GloVe embeddings for the generated words. **CIIC_O** and **CIIC_G**: The subscripts “O” and “G” represent that our CIIC is implemented with the one-hot word vectors and GloVe word em-

Table 1. Ablation experiments. All the models are trained with the XE loss. B@1, B@4, M, R, C, S, CHs and CHi are short for BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr, SPICE, CHAIR_s and CHAIR_i scores. “↓” and “↑” denote the lower the better and the higher the better, respectively.

| Model | B@1↑ | B@4↑ | M↑ | R↑ | C↑ | S↑ | CHs↓ | CHi↓ |
|-------------------|-------------|-------------|-------------|-------------|--------------|-------------|------------|------------|
| Base | 75.9 | 35.5 | 28.0 | 56.5 | 114.1 | 21.0 | 8.4 | 6.3 |
| Base+GloVe | 76.3 | 36.2 | 28.2 | 56.8 | 115.9 | 21.3 | 7.9 | 6.0 |
| Base+ITD | 76.1 | 36.2 | 28.1 | 56.7 | 116.0 | 21.2 | 7.7 | 6.0 |
| Base+ITD+GloVe | 76.5 | 36.5 | 28.4 | 57.0 | 117.1 | 21.3 | 6.9 | 5.4 |
| Base+UFT | 77.0 | 36.7 | 28.4 | 57.1 | 117.5 | 21.5 | 5.8 | 3.9 |
| Base+UFT+GloVe | 77.1 | 36.9 | 28.1 | 57.0 | 117.9 | 21.3 | 5.9 | 3.9 |
| CIIC _O | 77.3 | 37.0 | 28.3 | 57.4 | 118.3 | 21.3 | 5.6 | 3.9 |
| CIIC _G | 77.5 | 37.3 | 28.5 | 57.4 | 119.0 | 21.5 | 5.3 | 3.6 |

beddings, respectively.

Results and Analysis. Table 1 shows the ablation experiments on different encoding and decoding modules with the number of attention blocks $L = 6$. From Table 1, we can observe that GloVe and CI schemas respectively bring an improvement over Base. When they are both employed, the performance can be further enhanced, which indicates that both of them are beneficial. The performance of Base+UFT and Base+UFT+GloVe is superior to that of Base+GloVe+CI, which confirms the effectiveness of the extracted IOD features. After respectively incorporating the CI module into Base+UFT and Base+UFT+GloVe, the performance of both CIIC_O and CIIC_G can be further significantly boosted (from 117.5 CIDEr to 118.3 CIDEr and from 117.9 CIDEr to 119.0 CIDEr, respectively), which further confirms the utility of deconfounding the visual and linguistic confounders in the sentence generation.

4.3. Quantitative Analysis

Results on the Karpathy Test Splits. In Table 2, we compare our CIIC with the SOTA models on the offline COCO Karpathy test split, including SCST [34], Up-Down [2], RFNet [19], GCN-LSTM [46], SGAE [42], ORT [15], AoANet [16], \mathcal{M}^2 Transformer [8], Transformer+CATT [45] and X-Transformer [28]. SCST applies the RL-based reward, which is widely used in the following methods. Up-Down and RFNet utilize the visual attention mechanism. GCN-LSTM and SGAE employ scene graphs and graph convolution networks. ORT incorporates geometry information into the transformer. AoANet exploits the relevance of attention results via a gate guided by the context. \mathcal{M}^2 Transformer proposes a fully-connected architecture between the encoder and decoder layers. Transformer+CATT incorporates a novel causal attention into the Transformer architecture. X-Transformer applies Bilinear Pooling to the attention module of Transformer.

For fair comparisons, we conduct experiments to compare our proposed CIIC with the transformer-based methods on the same ResNext101 region-based features. From Ta-

ble 2, we can observe that CIIC_O achieves the best BLEU-1, BLEU-4 and ROUGE-L scores in comparison with the other SOTA methods. Particularly, it is even comparable to X-Transformer which is trained on 4 P40 GPUs ($4 \times 24 = 96G$)¹. CIIC_G surpasses other SOTA methods significantly in terms of BLEU-1, BLEU-4, CIDEr, ROUGE-L, while being competitive on METEOR and slightly worse on SPICE with respect to X-Transformer.

Table 2. Experimental results of different models on the MSCOCO “Karpathy” test split.

| Model | B@1 | B@4 | M | R | C | S |
|---------------------------------|-------------|-------------|-------------|-------------|--------------|-------------|
| SCST [34] | - | 34.2 | 26.7 | 55.7 | 114.0 | - |
| Up-Down [2] | 79.8 | 36.3 | 27.7 | 56.9 | 120.1 | 21.4 |
| RFNet [19] | 79.1 | 36.5 | 27.7 | 57.3 | 121.9 | 21.2 |
| GCN-LSTM [46] | 80.5 | 38.2 | 28.5 | 58.3 | 127.6 | 22.0 |
| SGAE [42] | 80.8 | 38.4 | 28.4 | 58.6 | 127.8 | 22.1 |
| ORT [15] | 80.5 | 38.6 | 28.7 | 58.4 | 128.3 | 22.6 |
| AoANet [16] | 80.2 | 38.9 | 29.2 | 58.8 | 129.8 | 22.4 |
| \mathcal{M}^2 Transformer [8] | 80.8 | 39.1 | 29.2 | 58.6 | 131.2 | 22.6 |
| Transformer+CATT [45] | - | 39.4 | 29.3 | 58.9 | 131.7 | 22.8 |
| X-Transformer [28] | 80.9 | 39.7 | 29.5 | 59.1 | 132.8 | 23.4 |
| CIIC _O | 81.4 | 40.2 | 29.3 | 59.2 | 132.6 | 23.2 |
| CIIC _G | 81.7 | 40.2 | 29.5 | 59.4 | 133.1 | 23.2 |

Table 3. Performance comparison with the SOTA methods in the setting of single model on the online MS-COCO test server, where c5/c40 means employing 5/40 ground-truth captions for testing.

| Model | B@4 | | M | | R | | C | |
|-----------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|--------------|
| | 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 |
| CAVP [48] | 37.9 | 69.0 | 28.1 | 37.0 | 58.2 | 73.1 | 121.6 | 123.8 |
| SGAE [42] | 37.8 | 68.7 | 28.1 | 37.0 | 58.2 | 73.1 | 122.7 | 125.5 |
| CNM [43] | 37.9 | 68.4 | 28.1 | 36.9 | 58.3 | 72.9 | 123.0 | 125.3 |
| VSUA [12] | 37.4 | 68.3 | 28.2 | 37.1 | 57.9 | 72.8 | 123.1 | 125.5 |
| AOA-DICv1.0 [44] | 38.8 | 70.5 | 28.8 | 38.2 | 58.6 | 73.9 | 126.2 | 128.4 |
| Transformer+CATT [45] | 38.8 | 70.6 | 28.9 | 38.2 | 58.7 | 73.9 | 126.3 | 128.8 |
| CIIC _O | 38.5 | 70.0 | 28.9 | 38.4 | 58.4 | 73.8 | 126.3 | 129.2 |
| CIIC _G | 38.5 | 70.1 | 29.1 | 38.4 | 58.6 | 74.0 | 126.4 | 129.2 |

Table 4. The bias analysis of different models on MSCOCO Karpathy split.

| Model | B@4↑ | M↑ | R↑ | C↑ | CHs↓ | CHi↓ |
|-----------------------|-------------|-------------|-------------|--------------|------------|------------|
| Up-Down [2] | 36.3 | 27.7 | 56.9 | 120.1 | 13.7 | 8.9 |
| Transformer | 38.4 | 28.6 | 58.4 | 128.6 | 12.1 | 8.1 |
| UD-DICv1.0 [44] | 39.0 | 28.8 | 58.8 | 128.8 | 10.1 | 6.5 |
| Transformer+CATT [45] | 39.4 | 29.3 | 58.9 | 131.7 | 9.7 | 6.5 |
| CIIC _O | 40.2 | 29.3 | 59.2 | 132.6 | 8.2 | 5.0 |
| CIIC _G | 40.2 | 29.5 | 59.4 | 133.1 | 7.7 | 4.5 |

Results on the Official Test Server. Table 3 reports the performance of different models on the online COCO test

¹<https://github.com/JDAI-CV/image-captioning/issues/7>

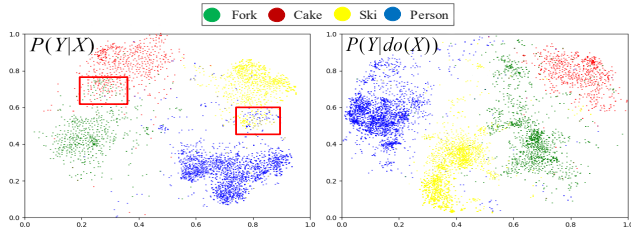


Figure 6. The t-SNE visualization [22] of some object features extracted by Faster R-CNN (left) and IOD (right).




| | | |
|-------------|--|---|
| Object Bias |  | Transformer: a person walking on a beach in the water. CIIC: a person walking on the beach next to the ocean. GT: a person on the beach next to the ocean. |
| |  | Transformer: a young boy eating a piece of cake. CIIC: a young boy eating a sandwich on a bench. GT: a young boy sitting on a bench with a sandwich. |
| Action Bias |  | Transformer: a young boy holding a tennis racket at a tennis ball. CIIC: a young boy hitting a tennis ball with a racket. GT: a person hitting a tennis ball with a tennis racket. |
| |  | Transformer: a car is sitting at a traffic light. CIIC: a car stopped at a traffic light on a street. GT: a car stopped at a traffic light on a city street. |
| Gender Bias |  | Transformer: a man and a dog on a paddle board in the water. CIIC: a man and a woman on a paddle board with a dog. GT: a woman riding a paddle board with their little dog. |
| |  | Transformer: a man riding on the back of a motorcycle. CIIC: a man and a woman sitting on a motorcycle. GT: a painted picture of a man and a woman on a motorcycle. |

Figure 7. Some generated captions by CIIC and the Transformer baseline in the case of gender, object and action biases. The green contexts denote the linguistic confounders which may induce biases. The correct and incorrect words are colored by blue and red, respectively.

server. For a fair comparison, we still train CIIC and competitive models in the same single model setting on the official test split. Compared with the top-performing approaches on the leaderboard, we can see that our single models still achieve the superior performance against the competitive methods. Particularly, CIIC_G can achieve a new state-of-the-art score of 126.4 on CIDEr (C5) and 129.2 on CIDEr (C40).

Analysis of the Biases. To confirm whether the proposed CIIC model can mitigate the dataset bias or not, we further evaluate the bias degree of the generated captions in Table 4. From Table 4, we can see that after performing causal intervention in object detection and image captioning, CIIC achieves the lowest CHs and CHi, which indicates that CIIC can generate the least biased captions. Meanwhile, we can see that CIIC obtains the best results in terms of BLEU-4, METEOR, ROUGE-L and CIDEr, which can further demonstrate that our CIIC model generates the more grounded captions in the case of dataset bias. Compared with Table 1, it can be found that CHs and CHi of

both CIIC_O and CIIC_G also increase with the increase of the CIDEr scores. This is due to the fact that the widely-adopted SCST optimization may cause biases in order to improve the CIDEr score [35].

4.4. Qualitative Analysis

Finally, we qualitatively evaluate the performance of our method. Figure 6 visualizes some visual features of MS-COCO images extracted by Faster R-CNN (left) and the proposed IOD (right). We can see that our IOD can learn the more discriminative feature representations compared to Faster R-CNN. For example, cake and fork features as well as person and ski features are entangled in red box when the conventional likelihood $P(Y|X)$ (left) is used. After causal intervention $P(Y|do(X))$ (right), they are clearly disentangled, implying that the IOD actually deconfounds the visual confounder while extracting the visual features. Figure 7 shows some captions of test images generated by CIIC and the Transformer baseline. Intuitively, CIIC is able to produce more grounded and less biased captions compared with the Transformer baseline. For example, our model effectively alleviates the spurious correlation between the boy feature and the word “cake” caused by both the visual and linguistic confounders. Besides, our CIIC can also alleviate gender and action biases, which indicates that our CIIC is able to effectively deconfound both the visual and linguistic confounders and further validates the effectiveness of our method.

5. Conclusion

In this paper, we present CIIC, a novel Transformer-based architecture for image captioning from the causal perspective, which seamlessly incorporates causal intervention into both object detection and captioning generation to jointly alleviate the confounding effect. On one hand, the proposed IOD effectively disentangles the visual features and facilitates the deconfounding of image captioning. On the other hand, the proposed ITD implements causal intervention to tackle the visual and linguistic confounders simultaneously during the generation of sentences. Experimental results have demonstrated that our method can significantly outperform the state-of-the-art image captioners in the single-model configuration on the MS-COCO dataset. The limitations of our method are given in the supplementary material.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.61403394, No.62172417 and No.62106268), and the High-Level Talent Program for Innovation and Entrepreneurship (Shuang Chuang Doctor) of Jiangsu Province (JSSCBS20211220).

References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Adaptive Behavior*, pages 382–398, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. [1](#), [2](#), [4](#), [6](#), [7](#), [12](#)
- [3] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5561–5570, 2018. [2](#)
- [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. [2](#)
- [5] Satantjeet Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *ACL-2005*, pages 228–231, 2005. [6](#)
- [6] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017. [2](#), [3](#), [4](#), [5](#)
- [7] Terry Copeck and Stan Szpakowicz. Text summarization branches out. *Association for Computational Linguistics*, 2004. [6](#)
- [8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. [1](#), [7](#)
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics, 2019. [2](#)
- [10] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017. [1](#), [6](#)
- [11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. [2](#)
- [12] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. Aligning linguistic words and visual semantic units for image captioning. *ACM*, 2019. [2](#), [7](#)
- [13] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020. [2](#)
- [14] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. [13](#)
- [15] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11135–11145, 2019. [1](#), [2](#), [7](#)
- [16] Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. [7](#), [12](#)
- [17] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1655–1663, 2021. [2](#)
- [18] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik G. Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. [1](#)
- [19] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *European Conference on Computer Vision*, 2018. [7](#)
- [20] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems 27*, pages 1889–1897, 2014. [6](#)
- [21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014. [6](#), [11](#)
- [22] Van Der Maaten Laurens and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2605):2579–2605, 2008. [8](#), [11](#)
- [23] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937, 2019. [2](#)
- [24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [2](#), [6](#)
- [25] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. [2](#)
- [26] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017. [2](#), [3](#), [4](#), [5](#)
- [27] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong

- Ji. Dual-level collaborative transformer for image captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 2286–2293. 2
- [28] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020. 1, 2, 7
- [29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Blue: A method for automatic evaluation of machine translation. In *Meeting of the Association for Computational Linguistics*, 2002. 6
- [30] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 3
- [31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014. 6
- [32] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020. 2, 4
- [33] 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. 1, 2, 4
- [34] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 6, 7
- [35] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018. 6, 8
- [36] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 1, 6
- [37] 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, 5, 6
- [38] 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 (CVPR)*, 2015. 6
- [39] 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. 1, 2
- [40] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 2, 3, 4, 5
- [41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1, 2, 4, 5, 11
- [42] 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. 1, 7
- [43] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4250–4260, 2019. 7
- [44] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 2, 3, 6, 7
- [45] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9847–9857, 2021. 3, 7
- [46] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 7
- [47] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 4
- [48] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *CoRR*, abs/1906.02365, 2019. 7
- [49] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *Advances in Neural Information Processing Systems 33*, 2020. 2
- [50] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2021. 2
- [51] Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. Counterfactual off-policy training for neural response generation. *arXiv preprint arXiv:2004.14507*, 2020. 2, 3