Dependent Multi-Task Learning with Causal Intervention for Image Captioning

Wenqing Chen\textsuperscript{1,2}, Jidong Tian\textsuperscript{1,2}, Caoyun Fan\textsuperscript{1,2}, Hao He\textsuperscript{1,2}\textsuperscript{*} and Yaohui Jin\textsuperscript{1,2}\textsuperscript{*}

\textsuperscript{1}MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
\textsuperscript{2}State Key Lab of Advanced Optical Communication System and Network, Shanghai Jiao Tong University

\{wenqingchen, frank92, fcy3649, hehao, jinyh\}@sjtu.edu.cn

Abstract

Recent work for image captioning mainly followed an extract-then-generate paradigm, pre-extracting a sequence of object-based features and then formulating image captioning as a single sequence-to-sequence task. Although promising, we observed two problems in generated captions: 1) content inconsistency where models would generate contradicting facts; 2) not informative enough where models would miss parts of important information. From a causal perspective, the reason is that models have captured spurious statistical correlations between visual features and certain expressions (e.g., visual features of “long hair” and “woman”). In this paper, we propose a dependent multi-task learning framework with the causal intervention (DMTCI). Firstly, we involve an intermediate task, bag-of-categories generation, before the final task, image captioning. The intermediate task would help the model better understand the visual features and thus alleviate the content inconsistency problem. Secondly, we apply Pearl’s do-calculus on the model, cutting off the link between the visual features and possible confounders and thus letting models focus on the causal visual features. Specifically, the high-frequency concept set is considered as the proxy confounders where the real confounders are inferred in the continuous space. Finally, we use a multi-agent reinforcement learning (MARL) strategy to enable end-to-end training and reduce the inter-task error accumulations. The extensive experiments show that our model outperforms the baseline models and achieves competitive performance with state-of-the-art models.

1 Introduction

Image captioning is a typical vision-to-language problem, which aims to automatically generate descriptions for images [Chen et al., 2015]. Recently, a line of studies used an extract-then-generate paradigm with a preliminary task, object detection [Anderson et al., 2018]. These studies first extracted

\textsuperscript{*}Corresponding authors

\textsuperscript{1}The mentioned image and the tested model can be found in Section 4.4 and Section 4.1, respectively.
icates. For instance, models are possible to capture the spurious statistical correlation between the visual features of "long hair" and "woman" through the confounder "female". Once "long hair" is detected, models may incorrectly predict "man with long hair" as "woman". To further explain this phenomenon, we show a typical causal graph in Figure 1. Let \( x \) and \( y \) be the given image and one ground-truth caption, respectively. We show the directed acyclic graph (DAG) including the mediator, the confounder, and the covariate. A robust model is supposed to utilize the mediator for prediction, which has causal effects on \( y \), rather than utilizing the dashed lines for prediction.

In this paper, we propose a dependent multi-task framework with the causal intervention (DMTCI) to eliminate the spurious correlations. Firstly, we formulate the image caption as a dependent multi-task (DMT) learning problem [Chen et al., 2020], which includes two tasks: bag-of-categories (BOC) generation and caption generation. The outputs of the first task are considered as the mediators, which are already annotated in certain datasets like MSCOCO and also utilized by recent work [Wang et al., 2020; Li et al., 2020]. Secondly, in terms of the confounders, we are probably to have no access to the real confounders. For example, "female" is not annotated. Fortunately, an available approach is using so-called "proxy confounders" [Louizos et al., 2017]. Since the spurious correlations problem we observe is that the models over-generating high-frequency categories or predicates, we let the proxy confounders be sampled from a set of concepts including high-frequency predicates and fine-grained object categories. Then we apply the variational inference [Kingma and Welling, 2014] to estimate the real confounders and let it sampled from a prior distribution. As a side advantage, our model can also infer covariates from the captions. Finally, we apply a multi-agent reinforcement learning strategy (MARL) to enable end-to-end training and reduce the inter-task error accumulations. An extensive set of experiments demonstrates that our model outperforms the baseline models and achieves competitive performance with those meticulously designed Transformers on the MSCOCO dataset. The main contributions of this work can be summarized as follows:

- We propose a dependent multi-task framework for image captioning, considering the BOC generation as an intermediate task to alleviate the inconsistency problem.
- We further introduce a latent de-confounding approach, which reduces the spurious correlations and estimates richer confounders through the variational inference.
- We cooperate with a multi-agent reinforcement learning approach to better utilize the task dependencies based on our multi-task framework and reduce the inter-task error accumulations.

2 Related Work

We categorize recent related work on image captioning into three groups according to the different focus of the work.

Architectural Innovation. The first category of work mainly focused on developing novel networks with powerful representation abilities. Some early studies encoded each image as a single feature vector through a pre-trained convolutional neural network (CNN), and then fed it to recurrent neural networks (RNNs) based on long short-term memory units [Vinyals et al., 2015; Donahue et al., 2015]. Afterward, Xu et al. [2015] explored the attention mechanisms for image captioning, which could dynamically focus on salient features at each decoding step. Another important line of work cooperated the task "object detection" for image captioning [Karpathy and Li, 2015; Fang et al., 2015]. Anderson et al. [2018] extracted object-level features by a pre-trained Faster R-CNN model, and combined them with the attention mechanism, which achieved significantly higher performance than object-free models. Follow-up research further captured the relationships among objects by adding self-attention on object-level features [Huang et al., 2019], or using graph convolutional networks [Yao et al., 2018; Yang et al., 2019]. More recently, Transformer-based [Vaswani et al., 2017] models were proposed with spatial position embeddings and better normalization layers [Guo et al., 2020], meshed attentions [Cornia et al., 2020] or higher-order attentions [Pan et al., 2020]. It is worthwhile to mention that almost all the state-of-the-art (SOTA) models utilized the self-critical reinforcement (SCRL) strategy [Rennie et al., 2017].

Knowledge Enhancement. The second category of work involved prior knowledge. Kim et al. [2019] proposed a relationship-based captioning model that considered part-of-speech (POS, i.e. subject-object-predicate categories) tags of captions, and therefore introduced beneficial inductive bias. Shi et al. further explored the semantics conveyed in the subject-object-predicate triples to enhance image caption generation [Shi et al., 2020]. The subject-object-predicate triples are extracted from captions by a textual scene graph parser introduced in [Schuster et al., 2015].

Causal Intervention. The third category of work is quite related to our work, which mainly focused on reducing the spurious correlations with causal intervention. Wang et al. used causal intervention to better learn the region features of the object detection task [Wang et al., 2020]. When trying to learn the conditional probability between two objects like "toilet" and "person", they assumed other objects as confounders and learned the objective of \( p(\text{toilet} | \text{do(person)}) \) by cutting off the link between other co-occurred objects and the object "person". Yang et al. assumed the dataset as the confounder and adjust it as concepts, such as "apple" and "green". Then they cut off the link between these concepts and image features.

3 Methodology

In this section, we first formulate the image captioning task as a dependent multi-task learning problem in Section 3.1. Then we illustrate our mediation and de-confounding approach in Section 3.2 and 3.3, with the model details in Section 3.4. Finally, we introduce the multi-agent reinforcement learning approach in Section 3.5.
3.1 Dependent Multi-Task Learning

Given the image and one corresponding ground-truth caption \((x, y)\), we use a pre-trained Faster R-CNN network [Anderson et al., 2018] to extract a sequence of features \(R \in \mathbb{R}^{L_r \times d}\) from \(x\) where \(L_r\) and \(d\) denote the number of the detected objects and the dimension of features, respectively. The pre-trained Faster R-CNN network is fixed during further training. Recent work mainly formulated image captioning as a single-task learning problem, directly maximizing \(p(y|R)\) while we formulate image captioning as a DMT problem, which maximizes the log-probability:

\[
\log p(y, m|R) = \log p(m|R) + \log p(y|R, m) \tag{1}
\]

where \(m\) is an intermediate variable conveying important information to generate final captions. We choose the BOC as \(m\) which are already annotated in MSCOCO dataset and also utilized in recent work [Wang et al., 2020; Li et al., 2020].

3.2 Mediation with DMT

There is a train-test discrepancy in Equation 1. In the training stage, we can use ground-truth object categories to supervise the learning objective \(\log p(m|R)\) while in the inference stage, \(m\) is not available. To reduce the train-test discrepancy, we include \(\bar{m}\), the predicted BOC, for further training after pre-training with Equation 1. The learning objective of caption generation becomes:

\[
p(y|R) = \sum_{m} p(y|R, m)p(m|R) \\
= \mathbb{E}_{\bar{m} \sim p(m|R)}[p(y|R, \bar{m})] \tag{2}
\]

where \(\mathbb{E}\) denotes the expectation operation. Then the multi-task objective is maximizing:

\[
\log p(m|R) + \log p(y|R) \\
= \log p(m|R) + \mathbb{E}_{\bar{m} \sim p(m|R)}[\log p(y|R, \bar{m})] \tag{3}
\]

which utilizes the dependency between tasks and meanwhile reduces the train-test discrepancy. In the period of maximum likelihood estimation (MLE), we use the Gumbel-softmax approach [Jang et al., 2017] when sampling \(\bar{m} \sim p(m|R)\) which can back-propagate gradients between tasks (details can be found in Appendix A in the longer version on Arxiv).

![Diagram](image)

Figure 2: Illustration of the change of causal graphs after the do-calculus, which include the mediator \(m\), the proxy confounder \(c\), and the latent confounder \(z_c\).

3.3 Latent De-Confounding

However, for the de-confounding causal graph, the causal intervention will cut off the arrows pointing to \(x\). We denote the proxy confounder and the real confounder by \(c\) and \(z_c\), respectively. We apply the do-calculus on \(x\) and obtain the de-confounding graph shown at the bottom of Figure 2. Firstly, we only consider \(c\) and obtain the probability distribution:

\[
p(y|do(R)) = \sum_{c} p(y|R, c)p(c|do(R)) \\
= \sum_{c} p(y|R, c)p(c) = \mathbb{E}_{\tilde{c} \sim p(c)}[p(y|R, \tilde{c})] \tag{4}
\]

which makes \(p(y|do(R)) \neq p(y|R)\) as \(p(c|do(R)) = p(c)\). It means that only considering the mediators \(m\) is not optimal during causal intervention. When jointly considering the mediator and the confounder, we do an intervention so that the mediator \(m\) is no longer affected by the confounder \(c\). Then we have:

\[
p(y|do(R)) = \sum_{c} \sum_{m} p(y|R, m, c)p(m|R)p(c) \\
= \mathbb{E}_{\tilde{c} \sim p(c), \bar{m} \sim p(m|R)}[p(y|R, \bar{m}, \tilde{c})] \tag{5}
\]

according to the mediation-confounding graph in Figure 2.

The next important thing is choosing the proxy confounder \(c\) and estimating the real confounder \(z_c\). Generally, confounders will have causal effects on \(x\) and \(y\). For example, "female" may have causal effects on the visual features of "long hair" in \(x\) and the word "woman" in \(y\). Those models modeling a high posterior probability \(q(female|long \ hair)\) may incorrectly take a man with long hair as a woman. In this case, "female" may be a workable confounder but not be annotated. Thanks to the so-called proxy confounder, we can estimate \(z_c\) from proxy confounders \(c\) (e.g., "girl"). Since we have observed that models over-generated high-frequency expressions, we assume that the proxy confounder \(c\) can be sampled from a concept set that includes the high-frequency predicates and fine-grained object categories. We use the summarized predicates and fine-grained categories in [Shi et al., 2020] as the concept set. The real confounders \(z_c\) has causal effects on \(c\) and we show the final causal graph of mediation and latent de-confounding in Figure 2.

Estimating \(z_c\) requires the variational inference [Kingma and Welling, 2014], where the probability distribution \(p(z_c)\)
is assumed to be priorly defined, such as a zero-mean unit-variance Gaussian distribution. Taking the modeling of the prior distribution $p(c)$ for example, we can use following approximation:

$$\log p_{\theta, \phi}(c) = \log \int_{z_c} p_{\theta}(c|z_c)p(z_c)dz_c$$

$$\geq \mathbb{E}_{z_c \sim q_\phi(z_c|c)}[\log p_{\theta}(c|z_c)] - KL(q_\phi(z_c|c)\|p(z_c))$$

(6)

where the integral is approximated by maximizing the evidence lower bound (ELBO), and $\theta$ and $\phi$ denote the parameters of two networks modeling the prior and the posterior distributions, respectively. And $KL(\cdot \| \cdot)$ denotes the Kullback-Leibler divergence between two distributions. Then, we modify Equation 5 when considering $z_c$ as the confounder:

$$\log p_{\theta, \phi}(y|do(R)) =$$

$$\mathbb{E}_{\tilde{c} \sim p(c), z_c \sim q_\phi(z_c|\tilde{c}), \tilde{m} \sim p_\phi(m|R)}[\log p_{\theta}(y|R, \tilde{m}, z_c, \tilde{c})]$$

(7)

where $\tilde{c} \sim p(c)$ means that we randomly sample concepts from the dataset and $z_c \sim q_\phi(z_c|\tilde{c})$ means that we can infer the actually mattered confounders in the latent space. It is worth noting that we are not designing a network to estimate the posterior distribution $q_\phi(z_c|y,c)$ because $y$ and $c$ are not strongly coupled. Instead, as introduced in Equation 7, we model $p(c)$ in the dataset rather than making a too strong assumption of $c$ paired with $y$.

The final multi-task learning objective is maximizing:

$$\log p_{\theta_1, \phi}(y|do(R)) + \log p_{\theta_2, \phi}(y|do(R)) + \log p_{\theta, \phi}(c)$$

(8)

where $\log p_{\theta_1, \phi}(y|do(R))$ and $\log p_{\theta_2, \phi}(c)$ are computed through Equation 7 and 6 respectively.

3.4 Model Details

To implement the final learning objective, we introduce our model in detail, which can be divided into three main parts: the hierarchical encoder, the BOC generator, and the caption generator.

The Hierarchical Encoder

Given the sequence of object features $R$, we use a hierarchical encoder for further encoding, which consists of multiple layers of multi-head attention network followed with "adding and normalization" layers:

$$R^{(i+1)} = \text{Add}\&\text{Norm}({\text{Attn}}(i)(R^{(i)}, R^{(i)}, R^{(i)}))$$

(9)

where $i$ denotes the layer number, $R^{(0)} = R$, and $\text{Attn}(\cdot)$ is the multi-head attention network introduced in [Vaswani et al., 2017]. We choose the outputs of an intermediate layer $R^{(i,m)}$ to predict the BOC $m$, and the outputs of the final layer $R^{(n)}$ to predict the caption $y$.

The BOC Generator

The first task is a category-level classification task. For each image, we convert $m$ to the one-hot label matrix $Y^m \in \mathbb{R}^{L_m \times N_m}$ where $L_m$ denotes the number of all categories, and $N_m$ denotes the maximum number of detected objects belongs to each category. We use another attention-based network for prediction:

$$H^c = \text{Attn}(Q^c, R^{(i,m)}, R^{(i,m)})$$

(10)

where $Q^c \in \mathbb{R}^{L_m \times d}$ denotes the category-specific query matrix in the attention layer and is trainable. Then we use a fully-connected neural network (FCNN) followed with the softmax function along the last dimension to generate the probabilities of object numbers $\hat{Y}^m$. Then the loss of the first task is computed through cross-entropy:

$$\mathcal{L}^m = -\sum_{j=1}^{L_m} Y^m_j \log \hat{y}^m_j$$

(11)

where $Y^m_j$ and $\hat{y}^m_j$ are two one-hot vectors representing the number of detected objects belong to the $j$-th category in $Y^m$ and $\hat{y}^m$, respectively. Maximizing the log probability $\log p_{\varphi}(m|R)$ is implemented by minimizing $\mathcal{L}^m$.

The Caption Generator

In the second task, we mainly introduce how to implement $p_{\theta}(y|R, m, z_c, \tilde{c})$ and $q_\phi(z_c|c)$. Firstly, we need to re-encode the predicted labels $\tilde{m}$. Taking the gold label $m$ for example, we involve another two sequences $w^m = \{w^m_j\}$ and $id^m = \{id^m_j\}$ representing the words and category indexes, respectively. For example, $w^m = \{\text{person, dining table, pizza, ...}\}$ and $id^m = \{1, 2, 3, ..., m\}$ marks the number of corresponding categories. We encode the words of each category in $w^m$ with an LSTM layer:

$$h^m_{j,t} = \text{LSTM}(OH(w^m_j W_o, h^m_{j,t-1})$$

(12)

where $OH(\cdot)$ denotes the one-hot operation, $w^m_j$ denotes the $t$-th token in the $j$-th category in $w^m$, and $W_o$ denotes the word embedding layer. Assuming $H^m = \{h^m_{j,T_w}\}$ containing the hidden states at final tokens (at the $T_w$-th step) and $m_j$ representing the occurrence numbers of $j$-th category in the BOC $m$, we fuse three kinds of embeddings as follows:

$$e^m_j = (h^m_{j,T_w}; \text{OH}(id^m_j W_o) \odot \sigma(OH(m_j) W_m)$$

(13)

where $\odot$ denotes the element-wise multiplication and $\sigma$ denotes the sigmoid function. $W_o, W_m$ denote parameters of the category embedding and number embedding layers, respectively. In this way, the final embedding matrix $E^m = \{e^m_j\} \in \mathbb{R}^{L_m \times d}$ is category-words- and number-aware. Then we use the max-pooling operation to obtain a vector $z_m$ representing the global information for decoding. It is worth noting that we do not further encode $E^m$ with position-aware encoders such as LSTM or attention networks, because the BOC does not contain the detailed information. For example, BOC can not indicate whether a person is hitting a ball or holding a ball. Therefore, we only apply the pooling operation and get $z_m$.

Secondly, we need to infer $z_c$ through the network $q_\phi(z_c|c)$. We extract the high-frequency predicates and fine-grained categories from $y$ and convert them to an unordered set $c$. We embed $c$ to $E^c \in \mathbb{R}^{b_c \times d}$ through the word embedding layer and pass it to a two-layer Transformer without the position embedding, and obtain a sequence of hidden states $H^c \in \mathbb{R}^{L_c \times d}$. Since $q_\phi(z_c|c)$ is often approximated to a Gaussian distribution, we use the mean pooling of $H^c$ followed with two FCNNs to get the corresponding mean and
variance vectors, and then $z_c$ is sampled from this Gaussian distribution.

Thirdly, we need to implement $p_\theta(y|R, \tilde{m}, z_c, c)$ where $\tilde{m} \sim p_\phi(m|R)$, $z_c \sim q_\phi(z_c|c)$ and $c \sim p(c)$. After randomly selecting $c$ from the dataset, we obtain a vector $z_c$ by sampling from $q_\phi(z_c|c)$. Since $\tilde{m}$ has been already encoded to $z_m$, we simply use a top-down attention LSTM network [Anderson et al., 2018] to decode each token:

$$L^g = -\sum_{t=1}^{T} \log \left( p_\theta \left( y_t \mid y_{<t}, R^{(n)}, z_m, z_c \right) \right)$$

where $y_t$ denote the $t$-th token of $y$, and $y_{<t}$ denotes the sequence of tokens preceding $y_t$. The latent variable $z_m$ and $z_c$ are concatenated with the hidden states at all steps. Note that our model is also compatible with Transformer-based decoders. But recent empirical work has shown that Transformer-based decoders require meticulous designs, such as meshed attention [Cornia et al., 2020] based decoders. But recent empirical work has shown that Transformer-based decoders require meticulous designs, such as meshed attention [Cornia et al., 2020] or high-order attention [Pan et al., 2020], and greatly increase the time complexity when optimized with reinforcement learning (RL).

3.5 Multi-Agent Reinforcement Learning

RL is commonly used in the task of image captioning to reduce the train-test discrepancy [Ranzato et al., 2016]. Since we formulate the image captioning as the dependent multi-task problem, the inter-task train-test discrepancy also needs to be tackled. During the MLE optimization period, Gumbel-softmax approximation is used but will improve the gradient variance because of the temperature parameter [Nie et al., 2019]. Therefore, we use a multi-agent reinforcement learning strategy to train the model end-to-end based on the SCRL strategy on the two tasks alternatively.

SCRL for the BOC Generator. The Monte-Carlo sampling is used to generate $\tilde{m}$ and the arg-max operation to generate the baseline $\overline{m}$. The gradient of the BOC generator can be expressed by:

$$\nabla_{\theta} L(\varphi) \approx -(r_1(\tilde{m}) - r_1(\overline{m})) \nabla_{\theta} \log p_c(\tilde{m} | R)$$ (15)

where $r_1(\overline{m})$ is the reward provided by both two tasks. After feeding $\tilde{m}$ to the caption generator, we obtain $\tilde{y}$ through greedy decoding and compute the reward $r_1(\overline{m})$ by:

$$r_1(\overline{m}) = \lambda_1 s_1(\overline{m}, m) + \lambda_2 s_2(\overline{m}, y)$$ (16)

where $\lambda_1$ and $\lambda_2$ are the hyperparameters to balance the two rewards, and $s_1(\cdot)$ denotes the score function to match two BOCs, and $s_2(\cdot)$ denotes the CIDEr-D score function [Vedantam et al., 2015].

SCRL for the Caption Generator. Fed with the baseline BOC $\overline{m}$ and a sampled latent confounder $z_c \sim p(z_c)$, the caption generator can obtain $\tilde{y}$ by the Monte-Carlo sampling and $\tilde{y}$ by the greedy decoding. We use the SCRL strategy with the gradients computed as follows:

$$\nabla_{\theta} L(\varphi) \approx -(r_2(\tilde{y}) - r_2(\tilde{y})) \nabla_{\theta} \log p_g(\tilde{y} | R, \overline{m}, z_c)$$ (17)

where $r_2(\tilde{y}) = s_2(\tilde{y}, y)$. Furthermore, we adopt the max-max framework [Xiao et al., 2020] which maximizes the rewards of the two agents alternatively.

4 Experiments

4.1 Datasets and Settings

Datasets. We experiment on the MSCOCO dataset, which is the most popular dataset for image captioning. The original dataset contains about 82,783 training images and 40,504 validation images. Following most of the previous work, we first evaluate our model on the Karpathy data split [Karpathy and Li, 2015] with 5,000 images for validation, 5,000 images for testing, and the rest for training. Then we evaluate our model on the on-line COCO test server. We set the maximum decoding length to 16 which equals the maximum length of 95% of captions. We keep the words that occur no less than 5 times, resulting in a vocabulary of 10,369 words.

Evaluation and Settings. We apply five standard metrics for evaluation: CIDEr-D [Vedantam et al., 2015], BLEU [Pan et al., 2020], METROT [Banerjee and Lavie, 2005], ROUGE [Lin, 2004], and SPICE [Anderson et al., 2016]. All the metrics are computed with the publicly released code. ²

Compared Models. We compare our models with several models including SCST [Rennie et al., 2017], Up-Down [Anderson et al., 2018], AoANet [Huang et al., 2019], Up-Down+VC [Wang et al., 2020], WeakVRD [Shi et al., 2020], M² Transformer [Cornia et al., 2020], and XLAN [Pan et al., 2020]. Among them, XLAN is the recent SOTA with an advanced architecture, and Up-Down+VC is related to our work which uses causal intervention to reduce spurious correlations between objects. Also, we compare with the baseline model, TransLSTM, consisting of a Transformer encoder and an up-down LSTM attention decoder. It is worth noting that one may find that a recent model OSCAR [Li et al., 2020] greatly outperforms the models mentioned above, but we want to clarify that OSCAR made two unfair settings: 1) using more vision-language datasets for pre-training, including Conceptual Captions, SBU captions, Flickr30k, and GQA [Li et al., 2020]; 2) utilizing the ground-truth rather than generated object categories in the inference phase.

Hyperparameters. We use the pre-trained Faster R-CNN model [Anderson et al., 2018] to represent each image as an adaptive sequence of feature vectors $R$ where the number of detected objects $n \in [10, 100]$. The number of encoder layers is 4 and we use the output of the second layer to predict BOC. The dimension of each feature vector $v_i$ is 2,048 and then projected to 1,024. The dimensions of the word embedding layer, attention layers, and the LSTM-based decoder are all set to 1,024. The learning rate is initialized to 0.0001 and decreased by half when the CIDEr-D score does not increase in 2 epochs, with the minimum learning rate set to $5 \cdot 10^{-6}$. The batch size is set to 50. The model is firstly optimized with MLE for 30 epochs (Gumbel sampling $\tilde{m}$ after 15 epochs), and then optimized with MARL for another 35 epochs. The special setting of our model is that we can sample different $z_c$ to generate multiple captions. Therefore, we train an image-text selector to select the best of 20 generated captions for each image (details in Appendix B in the longer version on Arxiv). ³

²http://cocodataset.org/
³https://github.com/tylin/coco-caption
### 4.2 Main Results

Table 1 presents the performance on the Karpathy test split of our model as well as the compared models. As shown, our model DMTCI achieves average CIDEr-D scores of 120.2 and 131.7 and outperforms the baseline TransLSTM by 2.9 and 3.0 points in MLE and RL periods, respectively. When compared with the related model with causal intervention, "Up-Down+VC", our model outperforms it by 1.2 points on the CIDEr-D score. Furthermore, when combining with the trained selector, our model outperforms TransLSTM by 6.4 and 3.3 points in MLE and RL periods, respectively. And our model achieves competitive results with the recent SOTA models including: 1) Pipeline, learning the two tasks sequentially; 2) DMTCI-M, only considering the mediator; 3) DMTCI-LDC, only considering the confounder. To avoid misleading, we denote our full model by DMTCI-Full. Table 2 presents the single-model performance on COCO online test server. Compared with recent SOTA models which have reported single-model performance, DMTCI achieves competitive performance and achieves the best CIDEr-D score with 40 reference captions.

### 4.3 Ablation Study

To analyze which mechanisms are driving the improvements, we present an ablation study in Table 3. We obtain ablated models including: 1) Pipeline, learning the two tasks sequentially; 2) DMTCI-M, only considering the mediator; 3) DMTCI-LDC, only considering the confounder. To avoid misleading, we denote our full model by DMTCI-Full. Table 3 shows the performance of these ablated models. We can see that Pipeline suffers from the train-test discrepancy and falls behind the baseline model TransLSTM. DMTCI-M increases the CIDEr-D score by 2.2 and 2.6 points over TransLSTM in MLE and RL periods, respectively. DMTCI-LDC outperforms TransLSTM while performing slightly worse than the DMTCI-M. When compared with the two ablated mod-
In this paper, we propose a dependent multi-task framework with causal intervention for image captioning. We first involve BOC as the mediator to help the model understand the pre-extracted features, then we propose a latent de-confounding approach to alleviate the spurious correlation problem. Furthermore, we apply a multi-agent reinforcement learning strategy to better utilize the task dependencies and enable end-to-end training. The experiments show that our model yields competitive results with recent SOTA models and the proposed framework could improve content consistency and informativeness.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China under Grant 2018YFC0830400, and Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

References


