Supplementary Material for "Interventional Video Grounding with Dual Contrastive Learning"

Guoshun Nan[1,*] Rui Qiao[1,*] Yao Xiao[2,†] Jun Liu [3,‡] Sicong Leng[1] Hao Zhang[4,5] Wei Lu[1,‡]

[1] StatNLP Research Group, Singapore University of Technology and Design, [2] Shanghai Jiao Tong University, China

[3] Information Systems Technology and Design, Singapore University of Technology and Design, Singapore

[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore

[5] Institute of High Performance Computing, A*STAR, Singapore

{guoshun_nan,rui_qiao}@sutd.edu.sg, 119033910058@sjtu.edu.cn, jun_liu@sutd.edu.sg

sicong_leng@mymail.sutd.edu.sg, zhang_hao@ihpc.a-star.edu.sg, luwei@sutd.edu.sg

Abstract

This supplementary material involves the details the are omitted in the main paper due to space limitation. 1) Section A: a detailed introduction of causal inference and structural causal model. 2) Section B: experimental settings and another case study.

A. Causal Inference

In statistics, the phrase "correlation does not imply causation" [1] indicates the causal relationships between two variables may not solely rely on an observed association or correlation between them. Machine learning models are driven by training data and the conventional likelihood-based methods tend to learn high correlations from training set rather than causalities. Causal inference [5, 6] is able to mitigate the above spurious correlations and estimate the true causality, i.e., the effect of treatment variable on an outcome variable, by introducing the operation of intervention. The two predominant causal inference frameworks are structural causal model (SCM) [6] and potential outcomes [6] which are fundamentally connected. We adopt SCM in our video grounding task as it is able to explicitly model the relationships among variables to obtain the direct causal effect, so that the unexpected high correlations between the query and video from the dataset can be properly eliminated by backdoor adjustment and *do*-calculus. We will dive into the details of SCM in the following section.

A.1. Structural Causal Model

Structural Causal Model (SCM), which is commonly expressed as a directed acyclic graph (DAG), describes relevant variables of world and how they interact with each other. The direction of edges in the DAG refers to the causal relationship between the two connected nodes. An SCM consists of a set of endogenous (V) and a set of exogenous (U) variables, which are related by a set of functions (F) that determine the values of variables based on the values of their parents. The variables U can be considered as environment or noise that have minor affect on the value of V from inexplicable causes. These variables U are often not shown in the causal diagram. The set of endogenous variables V models the fundamental relationships between the important factors and the corresponding SCM is usually constructed according to human knowledge. Even though V are endogenous, some variables from V can still be unobserved, because the SCM only defines variables' causal relationships but not the availability from data. As shown in Figure 1 (a), the variable Z is the common cause of the variable X and Y. We name such variable Z a confounder because not considering it prevents us from understanding the causal relationship and results in spurious correlation between X and Y. Furthermore, a confounder is called unobserved if its statistics is not available from the dataset. The conventional machine learning models are good at learning the correlations $P(Y|X)$ from the training data, while the confounder Z is overlooked, resulting in unaddressed spurious correlations solely based on X and Y. An SCM facilitates us to build the relations among these variables and explicitly consider the impact of Z. To estimate the true causal effect from X to Y, we can use the calculus of interventions and compute the distribution $P(Y|do(X))$, which responds to a new causal graph with incoming edges towards X ($Z \rightarrow X$) cut off, as shown in Figure 1 (b). Intuitively, an inter-

*Equally contributed

†Work done during internship at StatNLP group, SUTD

‡Corresponding Authors

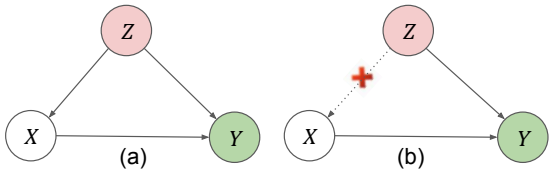

Figure 1: Structural Causal Model (SCM). (a) an SCM. The endogenous variables V consists of $\{X, Y, Z\}$. The variable Y is the child of the variable X and thus Y is directly caused by X. The variable Z is a confounder and both X and Y are the children of it. (b) An intervention on variable X is reflected in causal graph by cutting off the edge $Z \rightarrow X$.

vention is forcing the value of a variable, whereby relieves it from the control of its ancestral variables and consequently models the situation with no common cause between the variables of interest.

A.2. Backdoor Adjustment

To compute $P(Y|do(X))$, the spurious correlation caused by the confounder must be addressed. Backdoor adjustment is a technique that deconfounds the causal effect estimation. Following the goal of intervention, it requires an admissible set of variables that cut off every backdoor path from X to Y, which is defined as an unblocked trail from X to Y that contains an incoming edge towards X. A formal version of such requirement is called backdoor criterion [5]. As shown in 1 (a), Z is such admissible variable and the backdoor adjustment corresponding to Z is:

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

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

$$= \sum_z P(Y|X, z)P(z) \quad (3)$$

where $P(z|do(X)) = P(z)$ since X is now independent from Z due to intervention. The learning objective has now changed from $P(Y|X)$ to $\sum_z P(Y|X, z)P(z)$. The coufounding bias Z is now being explicitly marginalized out. Unfortunately, the latent confounder Z has no concrete form in video grounding. As we hypothesize them to be the selection bias in generating the dataset, a reasonable surrogate based on the actions that describe the video sequences is adopted.

B. Experiments

B.1. Settings

Table 1 shows the detailed hyper-parameters used in our experiments for the three datasets. We follow the previous studies Gao *et al*. [3], Yuan *et al*. [7], and Zhang *et al*. [8]

Hyper-Parameters	Value
Batch size	16
Learning rate	0.0005
Decay rate	0.01
Gradient clipping	1
Optimizer	Adam
Word embedding dimension	300
Char embedding dimension	50
Dropout	0.2
Kernel Size	7
Head	8
Hidden Dimension	128
Max Video Clips	128
α	0.1
β	0.01

Table 1: Hyper-Parameters for the three datasets.

to download or extract the pre-trained visual features and split the training, validation and test data. The video feature dimensions are 1024, 4096 and 500 for the Charades-STA, TACoS, and ActivityNet Caption datasets, respectively. For the Charades-STA, we follow the VLSNet [8] to extract visual features by the pre-trained *rgb_imagenet.pt*, which is provided by Joao *et al*. [2]. For the TACoS dataset, we use the pre-trained video features provided by Gao *et al*. [3]. For the ActivityNet Caption, we download the captions and pre-trained C3D visual features from its official websites [4].

B.2. Metrics

We follow the previous works [3, 7, 8] to use the "R@n, IoU = μ" and "mIoU" as the evaluation metrics. The metric "mIoU" refers to the average IoU over all testing samples. For the metric "R@n, IoU = μ", we follow the previous settings to configure $n = 1$ and $\mu = \{0.3; 0.5; 0.7\}$ for the Charades-STA and ActivityNet Caption, and $n = 1$ and $\mu = \{0.1, 0.3; 0.5; 0.7\}$ for the TACoS. It should be noted that we don't fine-tune a feature extractor in our experiments on the three datasets.

B.3. Case study

Figure 2 presents another case study to demonstrate the capability of our proposed `IVG-DCL` in alleviating the spurious correlations between text and video features. It shows that the previous model VSLNet [8] tends to locate the query #1 to the moment that relevant to the query #2, as the activities relevant to the "someone sits on the chair" are more commonly existed in the training set than the activities relevant to "someone stands on the chair". Our proposed `IVG-DCL` is able to find out more accurate moment

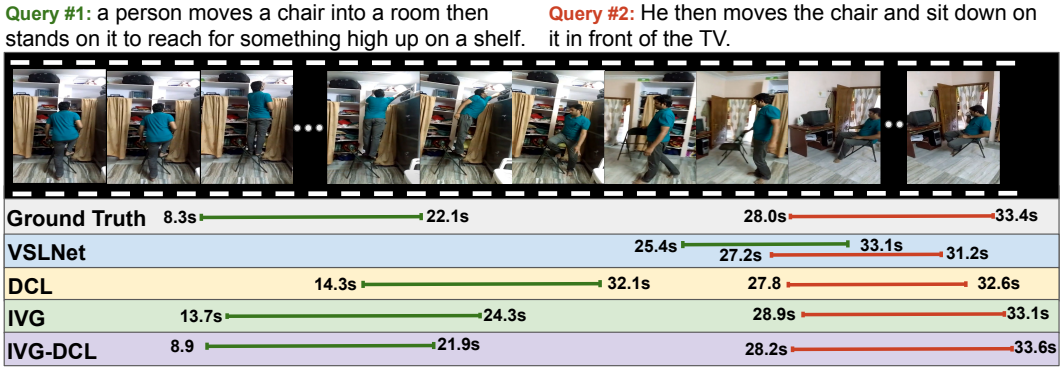

Figure 2: A case study on the Charades-STA dataset to demonstrate the capability of our model in mitigating the spurious correlations between textual and video features.

boundaries for both queries.

References

[1] John Aldrich et al. Correlations genuine and spurious in pearson and yule. *Statistical science*, 10(4):364–376, 1995. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[3] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *CVPR*, 2017. 2

[4] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *CVPR*, 2017. 2

[5] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 1, 2

[6] Donald B Rubin. Essential concepts of causal inference: a remarkable history and an intriguing future. *Biostatistics & Epidemiology*, 3(1):140–155, 2019. 1

[7] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 2

[8] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, 2020. 2