ERASURE FOR ADVANCING: DYNAMIC SELF-SUPERVISED LEARNING FOR COMMONSENSE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Commonsense question answering (QA) requires to mine the clues in the context to reason the answer to a question, and is a central task in natural language processing. Despite the advances of current pre-trained models, e.g. BERT, they often learn artifactual causality between the clues in context and the question because of similar but artifactual clues or highly frequent question-clue pairs in training data. To solve this issue, we propose a novel DynamIc Self-sUperviSed Erasure (DISUSE) which adaptively erases redundant and artifactual clues in the context and questions to learn and establish the correct corresponding pair relations between the questions and their clues. Specifically, DISUSE contains an *erasure sampler* and a *supervisor*. The erasure sampler estimates the correlation scores between all clues and the question in an attention manner, and then erases each clue (object in image or word in question and context) according to the probability which inversely depends on its correlation score. In this way, the redundant and artifactual clues to the current question are removed, while necessary and important clues are preserved. Then the supervisor evaluates current erasure performance by inspecting whether the erased sample and its corresponding vanilla sample have consistent answer prediction distribution, and supervises the KL divergence between these two answer prediction distributions to progressively improve erasure quality in a self-supervised manner. As a result, DISUSE can learn and establish more precise corresponding question-clue pairs, and thus gives more precise answers of new questions in present of their contexts via reasoning the key and correct corresponding clues to the questions. Extensive experiment results on the RC dataset (ReClor) and VQA datasets (GQA and VQA 2.0) demonstrate the superiority of our DISUSE over the state-of-the-arts.

1 INTRODUCTION

Given a context, e.g. an image that contains comprehensive logical relations among the objects, the commonsense question answering (QA) task aims at extracting the key clues, e.g. objects' relations or locations or properties (color, shape, etc), from the context to precisely reason the answer of a question. Because it has various real applications, such as AI customer service (Yoon et al., 2016), intelligent navigation (Das et al., 2018), and web-based QA system (Parthasarathy & Chen, 2007), the QA task has become one of the most important tasks in natural language processing and is widely studied in recent years.

One kind of popular and effective approaches for solving QA is pre-training methods. For example, transformer (Vaswani et al., 2017) based models, such as BERT (Devlin et al., 2019; Liu et al., 2019) and LXMERT (Tan & Bansal, 2019; Lu et al., 2019), are first trained from a large corpus, and then fine-tuned on labeled data of a specific downstream task. Despite the advance achieved by these pre-trained models, they still suffer from learning precise question-clue pairs to well reason the answer of the question from the clues in the context. Here clues is referred to as objects in an image for visual question answering(VQA) or words in a context for reading comprehension(RC). There are three possible reasons that results in the incorrect question-clue pairs in the pre-training methods. 1) The similar clues can confuse the model to well extract the exact clues for a specific question. For instance, Figure 1 (upper left) shows an example of this incorrect question-clue weakness. When

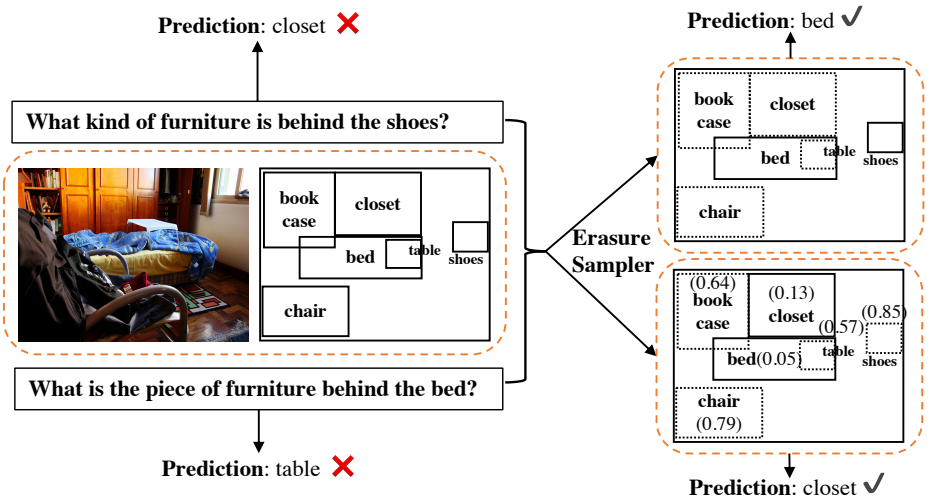

Figure 1: An overview of the DISUSE. The left part reveals the pre-trained models could learn artifactual causality for the question-clue pair relations, and make incorrect prediction. The right part reveals our DISUSE can relieve this faultiness by erasing less important clues and preserves more important ones in the form of probability under the help of erasure sampler. The dotted boxes represent the corresponding objects are erased, and the digits denote the probabilities of erasure.

asked the question that "What kind of furniture is behind the shoes?", the pre-trained models cannot learn the real positional information since the object "bed" and "closet" have a similar location when taking shoes as a reference in the 2D image. In this way, when querying the object near the shoes, it is hard to decide to choose bed or closet. 2) The highly frequent question-clue pair patterns can bias the learning of the model (Tang et al., 2020; Agarwal et al., 2020). Figure 1 (lower left) explains that when asked "What is the piece of furniture behind the bed?", the pre-trained models predict the erroneous "table" since the frequency of the "bed-table" pair is larger than the "bed-closet" pair in the training data. 3) For the abundant clues in the complicated context, it is also hard to find the key clues. This is because the words and objects pass message with each other through the self-attention mechanism, according to (Goyal et al., 2020), most of the clues in the context share information and present similar and redundant features in the pre-trained models.

Contributions. To resolve the above issue, we propose a novel and general DynamIc Self-sUperviSed Erasure (DISUSE) method which adaptively erases redundant and artifactual clues in the context and questions to learn and establish the correct corresponding question-clue pair relations. DISUSE is a general framework and can be applied to any pre-trained language models, e.g. BERT (Devlin et al., 2019) for RC, and multi-modal models, e.g. LXMERT (Tan & Bansal, 2019) for VQA. It consists of an *erasure sampler* and a *supervisor*. Given a context and a question, our erasure sampler first employs a self-attention layer at each layer of a pre-trained model to estimate 1) the correlation score of each clue in the context to the current question and 2) a soft boundary which determines how many clues will be erased. Then, based on the soft boundary, our sampler proposes a novel min-boundary normalization for transferring the correlation score of each clue into an erasure probability such that the clues with correlation scores lower than the soft boundary are considered as unnecessary and have higher probabilities of erasure. Finally, according to the estimated probability of erasure, our erasure sampler can erase the unnecessary and artifactual clues and preserves the key clues to the current question, leading to more precise corresponding question-clue pairs and better performance. For example, as shown in the right part of Figure 1, the objects of lower correlation scores are regarded as unnecessary and artifactual clues to the question and assigned to a higher probability of erasure. If the object "closet" is erased in Figure 1 (upper right), there will be less noisy information and thus the correct answer, namely, "bed", will have higher prediction confidence. If the object "table" are erased in the Figure 1 (lower right), the data bias can be alleviated to some extent. In both cases, erasing unimportant and artifactual clues can give higher change to obtain the precise clues and thus correct answers.

To evaluates the performance of the erasure sampler, the supervisor inspects whether the erased sample and its vanilla sample have consistent answer prediction distribution, and supervises the KL divergence between these two distributions to progressively improve erasure quality in a self-

supervised manner. Specifically, similar to self-supervised learning (Grill et al., 2020; Chen et al., 2020), DISUSE consists of online and target networks. It respectively feeds the vanilla context-question pairs and the erased pairs into online and target networks to predict the answer distribution $\boldsymbol{p_v}$ and $\boldsymbol{p_e}$ (probability of each answer option). If the current erasure sampler is of high quality, then the answer distribution $\boldsymbol{p_e}$ will be very similar to $\boldsymbol{p_v}$, since using the preserved key clues can still commendably predict the answer as same as all clues. In this case, $\boldsymbol{p_e}$ and $\boldsymbol{p_v}$ has small KL divergence. Otherwise, if $\boldsymbol{p_e}$ and $\boldsymbol{p_v}$ are far from each other and have large KL divergence, it means the quality of the preserved clues is poor since it cannot predict the answer as same as using all clues. Inspired by this key observation, the supervisor utilizes KL divergence to improve the erasure sampler by feeding back the evaluation information to it. The online and target networks are trained with cross-entropy loss over ground-truth answers as well as KL divergence and share parameters with each other following (Chen et al., 2020).

Experimental results on the RC dataset (Reclor (Yu et al., 2020)) and VQA datasets (Goyal et al., 2017; Hudson & Manning, 2019) demonstrate the advantages of the proposed method over pre-trained models. More importantly, our DISUSE significantly outperforms other contrastive methods by a large margin and achieving new state-of-the-art on all datasets. Furthermore, we also conduct extensive ablation experiments to validate the effectiveness and robustness of the proposed DISUSE.

2 Related Work

2.1 Commonsense Reasoning

Commonsense question answering (Yu et al., 2020; Hudson & Manning, 2019; Goyal et al., 2017) is a sub-branch of commonsense reasoning (Davis & Marcus, 2015) requires comprehensive understanding and compositional reasoning for the knowledge in the real world. Although this kind of knowledge and reasoning is natural to human beings, it is infamously disastrous for machines due to the lack of natural language inference ability (Storks et al., 2019). To address this issue, various approaches have been developed. (Kim et al., 2019) integrates attention mechanisms into densely connected RNN to promote the system to make accurate entailment and contradiction decisions. MAC network (Hudson & Manning, 2018) is proposed for VQA, consisting of chained cells each of which maintains a separation between control and memory. Another line of this research lies in building knowledge graphs. KG-MRC (Das et al., 2019) system is employed to construct dynamic knowledge graphs recurrently from procedural text. Bipartite knowledge graphs are generated to track the evolving states of entities. And ReGAT (Li et al., 2019) encodes RoIs of each image into a graph and then exploits explicit relations between objects and implicit relations between image regions to answer semantically-complicated questions grounded in an image.

2.2 Self-Supervised Learning

BERT-like models. Learning language representations from large-scale texts in an unsupervised manner has attracted extensive attention. Recently, there has been a large amount of pre-trained representation models, achieving dominating performance in commonsense reasoning tasks. We will choose several popular models to discuss in this section. Depending on the superiority of Transformer (Vaswani et al., 2017), unidirectional GPTs (Radford et al., 2018; 2019) which are trained with next-word prediction, and bidirectional BERT (Devlin et al., 2019) which regards masked language modeling (MLM) and next sentence prediction (NSP) as two pre-training tasks are proposed. Motivated by the success of BERT (Devlin et al., 2019) , other BERT-like models have been developed to further promote the performance of it. RoBERTa (Liu et al., 2019) is a robustly optimized method and trained with much larger mini-batches and learning rates to enhance on the MLM. Furthermore, cross-modal pre-trained models such as LXMERT (Tan & Bansal, 2019) and ViLBERT (Lu et al., 2019) are also presented for joint representations of vision and language. Since these methods heavily rely on artifactual causality between the clues in training data, we propose an adaptive method, which can erase redundant and unnecessary clues to establish the correct corresponding pair relations between the clues and their questions.

Self-supervised models. Recent unsupervised studies (Wu et al., 2018; Hjelm et al., 2019; Oord et al., 2018) are proposed for visual representation learning by employing contrastive loss (Hadsell et al., 2006). Since these methods build dynamic dictionaries, MoCo (He et al., 2020) maintains

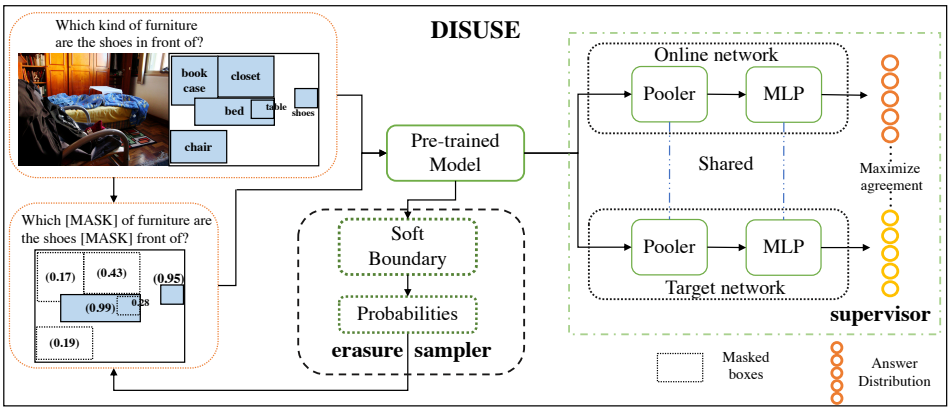

Figure 2: Framework of the proposed DISUSE for commonsense reasoning. The upper left part expresses the vanilla context/question pair and its erased pair generated by *erasure sampler* which can assign a probability of erasure to each clue and is shown in the middle part. The *supervisor* in the right part is employed to progressively improve the quality of erasure sampler by maximizing agreement between the erased sample and its vanilla sample over answer prediction distribution.

dictionaries that are large enough and consistent on-the-fly similar to BERT (Devlin et al., 2019). SimCLR (Chen et al., 2020)is introduced which requires neither specialized architectures (Hjelm et al., 2019) nor a memory bank (Wu et al., 2018), benefiting from the composition of data augmentation, learnable transformation, larger batch sizes, and more training steps. Inspired by the momentum update procedure in MoCo, BYOL (Grill et al., 2020) develops online and target networks to directly bootstrap the representation. Different from the existing contrastive loss upon feature space, we propose a new supervisor directly performed upon answer distributions.

3 METHOD

In this section, we elaborate on the proposed DISUSE for commonsense reasoning task. As illustrated in Figure 2, given a question $\mathcal{Q} = \{q_1, q_2, ..., q_m\}$ grounded in a context $\mathcal{C} = \{c_1, c_2, ..., c_n\}$, the goal of QA systems is to choose the most plausible correct answer $\hat{a}$ out of multiple answer options $\mathcal{A}$. For reading comprehension task, the context $\mathcal{C}$ denotes a paragraph and every $c_n \in \rfloor$ is a word. For visual question answering (VQA) task, the context $\mathcal{C}$ is a set of the objects detected by Faster R-CNN (Ren et al., 2015), in which each object is associated with its RoI feature and bounding-box feature (Tan & Bansal, 2019).

Suppose the network $p(\mathcal{A} \mid \mathcal{C}, \mathcal{Q}, \mathbf{\Theta})$ parametrized by $\mathbf{\Theta}$ can predict the probability (vector) of the options in $\mathcal{A}$ when feeding the context $\mathcal{C}$ and the question $\mathcal{Q}$. Based on these definitions, the vanilla training loss of the commonsense reasoning task can be formulated as

$$\mathcal{L}_{\text{QA}} = -\sum_{i=1}^{|\mathcal{A}|} \boldsymbol{y}_i \log \boldsymbol{p}_i \qquad \text{with } \boldsymbol{p} = p(\mathcal{A} \mid \mathcal{C}, \mathcal{Q}, \mathbf{\Theta}) \in \mathbb{R}^{|\mathcal{A}|}, \tag{1}$$

where the one-hot vector $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{|\mathcal{A}|}]$ denotes the ground-truth label of question $\mathcal{Q}$ in present of the context $\mathcal{C}$; $|\mathcal{A}|$ denotes the number of options in $\mathcal{A}$. After training the network $p(\mathcal{A} \mid \mathcal{C}, \mathcal{Q}, \mathbf{\Theta})$, one can easily predict the answer $\hat{a} = \text{argmax}_{a \in \mathcal{A}}\, p(\mathcal{A} \mid \mathcal{C}, \mathcal{Q}, \mathbf{\Theta})$.

However, since manually labeled data are expensive and are not sufficient in most piratical setting, the pure supervised learning model equation 1 cannot be well trained and usually suffers from unsatisfactory performance. To solve this issue, pre-training methods, e.g. BERT (Devlin et al., 2019; Liu et al., 2019) and LXMERT (Tan & Bansal, 2019; Lu et al., 2019), are developed. They first use a large corpus to train their model in an unsupervised manner, and then fine-tune the models on labeled data of a specific downstream task. But as aforementioned in Sec. 1, these methods cannot well learn and establish precise question-clue pairs to well reason the answer of a question. Since in absent of any guided information for the unsupervised pre-training, the similar clues, highly frequent question-clue pair patterns, and abundant clues in the complected context can easily confuse or bias the model to well extract the exact clues for a question (Goyal et al., 2020).

To solve this issue, we propose DISUSE which adaptively erases redundant and artifactual clues in the context and questions to learn and establish the correct question-clue pairs. As shown in Figure 2, DISUSE mainly contains an *erasure sampler* and a *supervisor*. DISUSE first concatenates a question $\mathcal{Q}$ and its context $\mathcal{C}$ into a pre-trained model, such as BERT, to generate hidden features and attention score matrics via self-attention mechanism (Vaswani et al., 2017) at each layer. For self-attention, one can compute the similarity between the query vector and key vector that both come from the same input as the attention scores which measures the importance of the query vector (a clue). Then the erasure sampler estimates the correlation scores between all clues and the question by using the generated hidden features and correlation scores, and then erases each clue according to the probability which inversely depends on its correlation score. Next, the supervisor is designed to inspect whether the erased sample and its vanilla sample have consistent answer prediction distribution and improve erasure quality in a self-supervised manner. In this way, DISUSE can learn more precise corresponding question-clue pairs and provides better reasoning performance. In the following, we will introduce the erasure sampler and supervisor in Sec. 3.1 and Sec. 3.2, respectively.

3.1 Erasure Sampler

Since the erasure sampler needs to adaptively erase redundant and artifactual clues in the context and questions, it is naturally critical to determine which clues are important and should be preserved, and which are unimportant and should be erased. Therefore, we propose a criterion called *soft boundary* to determine the importance of the clues in the context and question.

Soft Boundary. Our erasure sampler determines whether to erase each clue in the input of each layer in the pre-trained model. Here we employ a self-attention layer Att to implement our erasure sampler. Att takes embedding features of the context and question or hidden features denoted as $\hat{C}, \hat{Q}$ from last layer as input, and produces hidden features $\mathbf{F}$ and attention scores matrix $\mathbf{M}$: $\mathbf{F}, \mathbf{M} = \mathrm{Att}(\hat{\mathcal{C}}, \hat{\mathcal{Q}})$. The overall correlation scores of the i-th clue are then accumulated from attention scores: $\mathbf{Cor}_i = \sum_{i'} \mathbf{M}[i', i]$. Meanwhile, the hidden features $\mathbf{F}$ is fed to a MLP, a.k.a. projection head in self-supervised learning, to predict a probability $\boldsymbol{P}^C$ of each clue as $\boldsymbol{P}^C_i = 1 - \mathrm{Sigmoid}(\mathrm{MLP}(\mathbf{F}_i))$, where $\mathrm{Sigmoid}(\mathrm{MLP}(\mathbf{F}_i))$, linearly regressed from hidden features of each clue, measures the importance of it. Then we can determine the number of clues that should be erased as: $\mathrm{num} = \lfloor \mathbb{E}(\boldsymbol{P}^C) * \mathcal{N} \rfloor$, where $\mathcal{N}$ denotes the total clue number. Then the soft boundary is found via finding the num-th smallest value of correlation scores.

Probability of Erasure. Ignoring the correlation scores of clues, the aforementioned probability $\boldsymbol{P}^C$ is inadequate to determine which clues should be erased. We can conclude from the computing formula of correlation score $\mathbf{Cor}_i$: if some clues are essential, they will be attended by other clues and of high correlation scores. So the clue with correlation scores lower than the soft boundary is considered as unnecessary and artifactual clue, and is assigned to a higher probability of erasure. Thus the erasure sampler produces a fine probability $\boldsymbol{P}^F$ depending on correlation scores and soft boundary. Inspired by Min-Max normalization (Patro & Sahu, 2015), we implement a new algorithm called Min-Boundary normalization:

$$\boldsymbol{P}^F_i = 1 - \min\left(\max\left(\frac{\mathbf{Cor}_i - \mathbf{boundary}}{\mathbf{boundary} - \mathbf{Cor}_{\min}}, 0\right), 1\right),$$

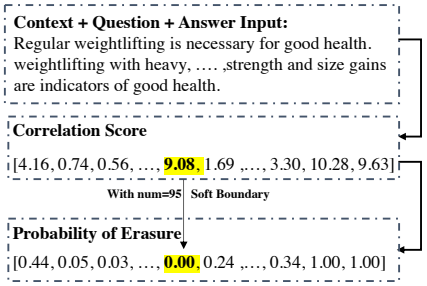

Figure 3: Detailed procedure of the erasure sampler. The highlight digits represent the soft boundary and its corresponding probability of erasure.

where $\mathbf{Cor}_{\min}$ indicates the minimum of correlation scores. Our erasure sampler benefits from this normalization in two aspects: 1) the sampler is able to generate personalized erased sample by assigning higher probabilities of erasure to unnecessary clues; 2) compared to Min-Max normalization which normalizes the value to $[0, 1]$, $\boldsymbol{P}^F_i$ will be 1 as long as its correlation score is no more than the soft boundary and $\boldsymbol{P}^F_i$ will be 0 as long as its correlation score is no less than $2 * \mathbf{boundary} - \mathbf{Cor}_{\min}$. Consequently, Min-Boundary normalization encourages the sampler to preserve more relatively important clues and erase more redundant ones.

Figure 3 shows an instance to how interpret the correlation score, soft boundary and probability of erasure are generated by the erasure sampler.

Hard Example Mining. The above strategy works well for easy and relatively hard samples which often dominate the whole datasets, since the attention layer could well estimate the attention score of each clues which is also demonstrated by our experimental results in Sec. 4. However, for very hard examples, the above erasure strategy cannot handle them, since the clues in these hard samples may be abundant and also similar and results in unreliable attention scores produced by self-attention mechanism. To solve this issue, we propose a random-explore erasure strategy. Specifically, when the confidence of ground-truth answers $\boldsymbol{p}^* \in \boldsymbol{p}$ predicted by the pre-trained model is smaller than the threshold $p_{\mathrm{S}} \in (0, 1)$ where $\boldsymbol{p} = p(\mathcal{A} \mid \mathcal{C}, \mathcal{Q}, \boldsymbol{\Theta}) \in \mathbb{R}^{|\mathcal{A}|}$ denotes the confidence of the options in set $\mathcal{A}$, we judge the question $\mathcal{Q}$ in present of the context $\mathcal{C}$ is a very hard problem. Please refer to the setting of p_{S} in Sec. 4. Then for these hard samples, we erasure each clue according to the soft boundary and the independent erased probability $\boldsymbol{P}^C$ instead of $\boldsymbol{P}^F$ in equation 3.1. In this way, the hard samples can be erased more randomly than easy ones. Once the erased samples preserve reasonable clues, then the supervisor can detect this correct erasure by inspecting whether the erased sample and its vanilla sample have consistent answer prediction distribution, and supervise and encourage the pre-trained model with attention layers to learn this kind of correct question-clue pairs. See more details of supervisor in the following subsection. As a result, DISUSE can gradually explore and learn correct question-clue pairs to improve the overall performance.

Since the experimental results shows that $\boldsymbol{P}^F$ leads to a large amount of clues to be erased, we perform linear regression as follows (Montgomery et al.) to push $\boldsymbol{P}^F$ close to a threshold thres:

$$\mathcal{L}_{\mathrm{Reg}} = -\sum_{i=1}^{\mathcal{N}} \mathbb{1}_{[\boldsymbol{p}_i^* < p_{\mathrm{S}}]} \mathrm{thres} * \log(\boldsymbol{P}_i^C) - (1 - \mathrm{thres}) * \log(1 - \boldsymbol{P}_i^C), \quad (0 < \mathrm{thres} < 0.5) \tag{2}$$

where $\mathcal{N}$ denotes the total clue number, $\mathbb{1}_{[\boldsymbol{p}^* < p_{\mathrm{S}}]} \in \{0, 1\}$ expresses an indicator function evaluating to 1 iff $\boldsymbol{p}^* < p_{\mathrm{S}}$ and thres indicates a threshold to limit $\boldsymbol{P}^C$.

3.2 Supervisor

After erasure, the supervisor inspects whether the erased sample and its vanilla sample have consistent answer prediction distribution. Specifically, DISUSE feeds the vanilla context-question pairs and the erased pairs into online and target networks to predict the answer distribution $\boldsymbol{p_v}$ and $\boldsymbol{p_e}$ (probability of each answer option), respectively. If the current erasure sampler is of high quality, then the answer distribution $\boldsymbol{p_e}$ will be very similar to $\boldsymbol{p_v}$, since using the preserved key clues can still commendably predict the answer as same as all clues. Otherwise, if $\boldsymbol{p_e}$ and $\boldsymbol{p_v}$ are far from each other, it means the quality of the preserved clues is poor since it cannot predict the answer as same as using all clues. Inspired by this key observation, the supervisor measures the quality of erasure by computing the KL divergence between $\boldsymbol{p_e}$ and $\boldsymbol{p_v}$:

$$\mathcal{L}_{\mathrm{Self}} = -\sum_{i=1}^{|\mathcal{A}|} \boldsymbol{p_e^i} \log(\boldsymbol{p_v^i}) \qquad \text{with } \boldsymbol{p_v^i} = \frac{\exp(\boldsymbol{z_v^i}/\tau)}{\sum_j \exp(\boldsymbol{z_v^j}/\tau)}, \quad \boldsymbol{p_e^i} = \frac{\exp(\boldsymbol{z_e^i}/\tau)}{\sum_j \exp(\boldsymbol{z_e^j}/\tau)}, \tag{3}$$

where $\boldsymbol{z_v}$ and $\boldsymbol{z_e}$ denote the logit produced by online and target networks, $\boldsymbol{p_v}$ and $\boldsymbol{p_e}$ represent the answer prediction distributions, and τ is a temperature. Here $\boldsymbol{p_e^i}$ denotes the i-th entry in $\boldsymbol{p_e}$. In this way, our supervisor can minimize KL divergence to improve the erasure sampler.

Now the overall objective can be written as a weighted average of two different objective functions:

$$\mathcal{L} = \alpha \mathcal{L}_{\mathrm{QA}} + (1 - \alpha)\tau^2 \mathcal{L}_{\mathrm{Self}} + \beta \mathcal{L}_{\mathrm{Reg}}, \tag{4}$$

where α and β are two constant to trade off these losses. The cross-entropy loss $\mathcal{L}_{\mathrm{QA}}$ in Eq. 1 is trained via supervised data which can well train the online network to predict the correct answer prediction distribution $\boldsymbol{p}_v$, while the self-supervised term $\mathcal{L}_{\mathrm{Self}}$ improves the quality of the learnt question-clue pairs by erasing redundant and artifactual clues in the context and questions and making consistent answer prediction distributions between erased sample and its vanilla sample. So these two terms are complement to each other. Superior to other contrastive loss (He et al., 2020; Grill et al., 2020) supervised upon feature space, our supervisor takes advantage of higher-level supervisory information e.g. answer prediction distributions, and directly evaluates the sampler to progressively improve the quality of erasure.

Table 1: Accuracy (%) on ReClor. The top part shows the comparison between pre-trained models (BERT, RoBERTa) and DISUSE. The bottom part shows the comparison between self-supervised methods (MoCo, SimCLR, BYOL) and DISUSE, where all methods are based on RoBERTa.

Method	Test Split		
	Overall Test	Test-EASY	Test-HARD
$\text{BERT}_{\text{base}}$ (Devlin et al., 2019)	47.3	71.6	28.2
DISUSE with $\text{BERT}_{\text{base}}$	49.5	75.2	29.3
$\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019)	48.5	71.1	30.7
MoCo (He et al., 2020)	50.3	75.5	30.5
SimCLR (Chen et al., 2020)	48.5	72.7	29.5
BYOL (Grill et al., 2020)	49.6	74.1	30.4
DISUSE	**52.7**	**78.9**	**32.1**
DISUSE w/o sampler	50.7	77.3	29.8
DISUSE w/o supervisor	48.7	72.5	30.0
DISUSE w/o regularizer $\mathcal{L}_{\text{Reg}}$	51.4	78.2	30.4

4 Experiments

4.1 Dataets

We evaluate the proposed DISUSE on three commonsense QA datasets: ReClor (Yu et al., 2020), VQA2.0 (Goyal et al., 2017) and GQA (Hudson & Manning, 2019). ReClor is a challenging reading comprehension dataset which extracts from 239 standardized graduate admission examinations. To facilitate comprehensive evaluation in the testing set, the data points with biases is grouped as EASY set, with the rest as HARD set. VQA 2.0 (Goyal et al., 2017) dataset consists of real images from MSCOCO (Lin et al., 2014). Each image is associated with 3 questions drawn from 3 categories, i.e. 1) $\mathrm{Y/N}$, 2) Number, and 3) Other answered by human annotators. GQA (Hudson & Manning, 2019) focuses on real-world visual reasoning and compositional question answering. The images of the training set come from Visual genome (Krishna et al., 2017) and those of the testing set come from MSCOCO (Lin et al., 2014). The questions are generated and visually grounded in the image by leveraging Visual Genome Scene Graphs (Krishna et al., 2017).

4.2 Experimental Setup

We implement DISUSE with Pytorch and conduct all experiments on a GeForce RTX 2080 GPU. For ReClor dataset, we first fine-tune the pre-trained models e.g. RoBERTa for 10 epochs with an initial learning rate of 1×10^{-5} and batch size of 8, and then re-train it by DISUSE for 8 epochs with a batch size of 3. We set $\alpha = 0.1$, $\beta = 1 \times 10^{-5}$, $\tau = 20$ in Eq. 4. Other implementation details such as optimizer, warm-up strategy, and maximum input sequence length are the same as (Yu et al., 2020). For VQA dataset, the top 5 of $\boldsymbol{p_v}$ and corresponding $\boldsymbol{p'_e}$, and the top 5 of $\boldsymbol{p_e}$ and corresponding $\boldsymbol{p'_v}$ are fed into Eq. 3. For this dataset, we set both weights for $\mathcal{L}_{\text{QA}}$ and $\mathcal{L}_{\text{Self}}$ as one. Given pre-trained LXMERT, we directly fine-tune DISUSE for 4 epochs without revising hyper-parameters. In addition, we modify some settings of the objective function in Eq. 4 for VQA 2.0 dataset. Due to minimizing the binary cross-entropy function (Li et al., 2019; Tan & Bansal, 2019), the KL divergence is replaced with binary cross-entropy. The threshold for important clue is $p_s = 0.5$ in all experiments. *Code is submitted in the supplementary and will be released online.*

4.3 Experimental Results

Comparison to State-of-the-arts. We report the experimental results in Tables 1, 2, and 3. As shown in Table 1, our DISUSE makes about 2.4% overall improvement on the ReClor dataset. For easy (Test-Easy) and hard (Test-H) samples in ReClor, DISUSE boosts the performance of RoBERTa by about 7.8% and 1.4%, respectively. Furthermore, to demonstrate the generalizability of our DISUSE, we conduct experiments on two VQA datasets, where the pre-trained LXMERT (Tan & Bansal, 2019) is regarded as the baseline, since the images of the training and testing splits are collected from different sources. Tables 2 and 3 show that apart from beating the multi-modality reasoning methods, our DISUSE enhance the pre-trained LXMERT on almost all evaluation metrics.

Table 2: Accuracy (%) on Test-Dev and Test-Standard splits of VQA 2.0 dataset. The top part represents traditional multi-modality reasoning methods. The middle part consists of pre-trained models and the bottom part is the comparison between self-supervised methods and DISUSE.

Method	Test-Dev	Test-Standard			
		Binary	Number	Other	**Accuracy**
BUTD (Anderson et al., 2018)	65.3	81.8	44.2	56.1	65.7
ReGAT (Li et al., 2019)	70.3	-	-	-	70.6
CMR (Zheng et al., 2020)	72.6	88.1	54.7	63.2	72.6
ViLBERT (Lu et al., 2019)	72.2	87.9	54.8	62.6	72.5
LXMERT (Tan & Bansal, 2019)	72.4	88.0	**54.9**	63.1	72.5
MoCo (He et al., 2020)	72.5	88.3	54.4	63.2	72.7
BYOL (Grill et al., 2020)	72.2	87.8	54.1	63.0	72.3
SimCLR (Chen et al., 2020)	72.5	88.3	54.6	63.2	72.7
DISUSE	**72.7**	**88.4**	54.3	**63.5**	**72.8**

Table 3: Accuracy (%) on the Test-Dev and Test-Standard splits of GQA.

Method	Test-Dev	Test-Standard						
		Binary	Open	Con.	Pla.	Val.	Dis.	**Acc.**
BUTD (Anderson et al., 2018)	-	66.6	34.8	78.7	84.6	96.2	6.0	49.7
MAC (Hudson & Manning, 2018)	-	71.2	38.9	81.6	96.2	84.5	5.3	54.1
LXMERT (Tan & Bansal, 2019)	60.0	77.2	45.5	89.6	84.5	**96.4**	5.7	60.3
MoCo (He et al., 2020)	59.6	77.1	45.1	89.4	84.4	96.2	5.3	60.1
BYOL (Grill et al., 2020)	59.8	76.9	45.1	89.4	84.3	96.2	**5.0**	60.0
SimCLR (Chen et al., 2020)	59.6	77.1	45.0	89.4	84.4	96.2	5.3	60.1
DISUSE	**60.5**	**77.4**	**46.3**	**90.9**	**84.9**	96.3	5.4	**60.9**

We also apply recent self-supervised learning frameworks, e.g. MoCo (He et al., 2020), SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020), into pre-trained models, and randomly mask clues like BERT to implement the comparison. Experimental results on the three datasets well demonstrate that our DISUSE outperforms them. Besides, we find that the self-supervised learning methods are inferior to the pre-trained models, because they attenuate the performance of RoBERTa in the Test-HARD splits of the ReClor dataset and GQA dataset. The possible reasons lie in two aspects: 1) the random masking procedure may make the models blind to essential clues; 2) the contrastive loss performed upon feature space is insufficient to supervise effectively for this task.

Ablation Study. To demonstrate the benefits of each component in DISUSE, we independently remove each component, i.e. erasure sampler or supervisor and evaluate the new models on ReColor. The bottom part of Table 1 reports the performances. By removing erasure sampler and using random clue masks, the accuracy degrades by 2.0%, 1.6% and 2.3% in the three splits. This testifies that the erasure sampler plays an essential role in erasing redundant and unimportant clues. Similarly, when replacing the KL measure in answer prediction distribution space of erasure sampler with InfoNCE loss in (Oord et al., 2018; He et al., 2020; Chen et al., 2020) which encourages feature similarity of the erased sample and vanilla sample, the accuracy of DISUSE also becomes worse, which shows the effectiveness and superiority of our supervisor. Table 1 also shows the effectiveness of the regularizer $\mathcal{L}_{\text{Reg}}$.

5 Conclusion

In this paper, we propose a novel DynamIc Self-sUperviSed Erasure (DISUSE) for commonsense reasoning. DISUSE designs an erasure sampler and a supervisor which respectively erases redundant clues in the context and questions and supervise the erasure quality in self-supervised manner. In this way, DISUSE can learn more precise corresponding question-clue pairs, and thus achieves better performance. Extensive experimental results demonstrate the superiority of our DISUSE.

REFERENCES

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR Workshops*, 2018.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. Building dynamic knowledge graphs from text using machine reading comprehension. In *ICLR*, 2019.

Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 2015.

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

Saurabh Goyal, Anamitra Roy Choudhury, Saurabh M Raje, Venkatesan T Chakaravarthy, Yogish Sabharwal, and Ashish Verma. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *ICML*, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2019.

Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. 2018.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Seonhoon Kim, Inho Kang, and Nojun Kwak. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *AAAI*, 2019.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, 2019.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neurips*, 2019.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sangeetha Parthasarathy and Jinlin Chen. A web-based question answering system for effective e-learning. In *ICALT 2007*, 2007.

S Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Neurips*, 2015.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.

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

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

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

Seunghyun Yoon, Mohan Sundar, Abhishek Gupta, and Kyomin Jung. Automatic question answering system for consumer products. In *IntelliSys*, 2016.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *ICLR*, 2020.

Chen Zheng, Quan Guo, and Parisa Kordjamshidi. Cross-modality relevance for reasoning on language and vision. 2020.