Inter-subtask Consistent Representation Learning for Visual Commonsense Reasoning

Kexin Liu, Shaojuan Wu, Xiaowang Zhang(✉), and Song Wang

College of Intelligence and Computing, Tianjin University, Tianjin, China
{kexinliu2020,shaojuanwu,xiaowangzhang}@tju.edu.cn, songwang@cec.sc.edu

Abstract. Given an image and a related question, Visual Commonsense Reasoning (VCR) requires to select the correct answer (question answering subtask) and the rationale (answer justification subtask). The commonsense semantic hidden between subtasks is essential for complex reasoning for VCR. Most of the previous studies of VCR focus on the leaning of cross-modal semantics and optimize the two subtasks independently, ignoring the hidden semantic correlations between answers and rationales. In this paper, we propose an Inter-subtask Consistent Representation Learning (ICRL) framework to learn the hidden commonsense semantics. Specifically, we design a joint learning framework to establish the connection between the two subtasks. Furthermore, we propose a multi-level contrastive learning network to ensure the semantic consistency of subtasks in the feature space. Experiments on the VCR dataset demonstrate that the proposed ICRL brings significant performance gain over the state-of-the-arts.

Keywords: VCR · Joint learning · Contrastive learning

1 Introduction

Given an image, the goal of VCR [23] is to answer a relevant question, and then provide a rationale justifying the answer. As show in Fig. 1, for **question answering subtask** ($Q \rightarrow A$), the model requires to answer a challenging question "*Why is [person1] holding a flower*" according to the given image correctly. Then the **answer justification subtask** ($QA \rightarrow R$) performs high-order understanding to explain why the answer "*[person1] is in the garden working*" is true. VCR is designed to perform complex cognitive-level reasoning, not just cognitive-level perception. A major challenge of VCR is to perform complex reasoning by learning hidden commonsense rather than surface clues. Inter-subtask consistency [22] is crucial for the learning of such hidden commonsense for VCR. For example, as shown in Fig. 1, the prediction results of the two subtasks are different aspects of one hidden commonsense "*People usually wear apron and hold a basket to work in the garden*". Hence inconsistency in the predictions implies contradiction (i.e., at least one of the predictions is wrong).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13531, pp. 149–161, 2022.
https://doi.org/10.1007/978-3-031-15934-3_13

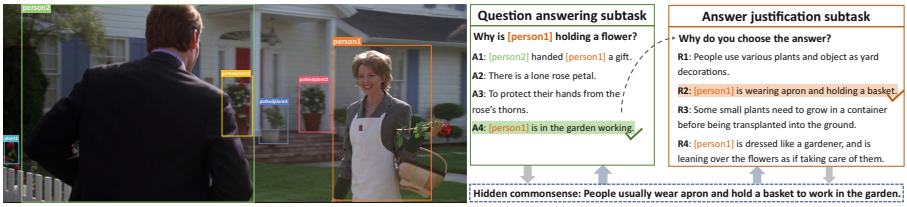

Fig. 1. An example of the VCR task. VCR requires the model to answer a question about an image while giving a rationale to justify the answer. Question answering and answer justification usually require a hidden commonsense.

Existing methods for the VCR can be broadly divided into two types: 1) utilize attention mechanisms to achieve semantic alignment across modalities for visual contents and linguistic expression understanding [19,21,23]. Such methods are often difficult to perform complex reasoning due to their inability to learn commonsense. 2) Introduce commonsense from external knowledge bases or out-of-domain datasets into models to help reasoning [5,15,17,18]. Despite achieving impressive results, these kind of methods usually suffer from the high cost of manual annotation of knowledge bases and are difficult to be implemented efficiently on hardware devices. All of the above methods perform reasoning and prediction on each individual subtask, ignoring the inter-subtask consistency of VCR and thus make it difficult for effective reasoning.

Inspired by the above discussion, we propose an Inter-subtask Consistent Representation Learning (ICRL) framework to learn hidden commonsense semantics for complex reasoning of VCR. Specifically, we design a joint learning framework based on pseudo-siamese network [3], which trains two subtasks in one model to achieve an interaction between two subtasks. Moreover, inspired by the achievements of contrastive learning [7] in representation learning, we propose a multi-level contrastive learning network, which contains answer-rationale contrastive learning (Answer-rationale CL) and visual-textual contrastive learning (Visual-textual CL) to learn consistent representations. Specifically, Answer-rationale CL pulls together correct answer and correct rationale representations while pushing apart mismatched response representations to ensure the semantic consistency of answers and rationales in the feature space. The Visual-textual CL pulls together visual and correct response representations while pushing apart visual and wrong responses representations to achieve cross-modal semantic alignment. In this way, we can learn latent inter-subtask hidden commonsense semantics to help the reason for both VCR subtasks.

The major contributions of this work are summarized as follows:

- We present a novel joint learning framework, which can jointly optimize question answering subtask and answer justification subtask of VCR to achieve interaction between answers and rationales.
- We propose to utilize multi-level contrastive learning (i.e., Answer-rationale CL and Visual-textual CL) for cross-modal semantic and inter-subtask consistent representation learning to improve the reasoning ability of VCR model.

- We conduct extensive experiments on the VCR dataset with comprehensive analysis and the results demonstrate that our model achieves a significant improvement in performance compared to state-of-the-arts.

2 Related Work

This section reviews the related work on VCR, Siamese Network, and Contrastive Learning.

2.1 Visual Commonsense Reasoning

The objective of VCR is to answer commonsense visual questions, and provide rationales justifying its answers. In order to give the correct answer and rationale, the VCR model requires not only a comprehensive understanding of visual scenes and language expressions, but also commonsense knowledge about how the world works. Some work has been proposed to address the challenges of VCR[13,19,21,23,24]. Rowan et al. [23] proposed an attention-based benchmark model R2C to obtain contextualized representations by performing an attention mechanism on visual and textual representations. It is worth noting that in the benchmark of Rowan et al. [23], the two subtasks of question answering and answer justification use the same model structure but are trained separately thus have no direct influence on each other. Later methods also followed this approach to handle the two subtasks. Weijiang et al. [21] constructed question-answer heterogeneous graph and vision-answer heterogeneous graph to achieve semantic alignment among vision, questions and answers. Zhang et al. [24] proposed multi-level counterfactual contrastive learning framework(MCC) for VCR to learn inter-modal and intra-modal representations. In order to combine commonsense knowledge to solve VCR task, some works introduce external knowledge into the model to help reasoning [12,15,18]. Dandan et al. [15] extracted knowledge from knowledge graph to learn knowledge combined representations. Zhang et al. [18] proposed to transfer external knowledge into visual content through transfer learning. Different from all of these methods, our proposed ICRL jointly optimize question answering subtask and answer justification subtask, and learn commonsense knowledge from the interaction of answers and rationales for VCR.

2.2 Siamese Network

Siamese network is a kind of neural structure containing two or more identical structures (e.g., answers encoder and rationales encoder in Fig. 2) to make multi-class prediction or entity comparison [3]. Siamese network is implemented by sharing the parameters of the neural networks. The applications of Siamese network include signature, object tracking, image matching and others. In this work, we adopt siamese network to achieve interaction and information sharing between question answering subtask and answer justification subtask for VCR.

2.3 Contrastive Learning

Contrastive Learning (CL) [7] is a commonly used self-supervised learning method. Typical contrastive learning works learn representations by a contrastive loss which pushes apart dissimilar samples while pulling together similar samples. Formally, given input sample q, contrastive learning methods sample similar samples to produce positive pairs (q, q^+) and dissimilar samples to produce negative pairs (q, q^-). Common contrastive loss function has the general form:

$$L = -\log \frac{\exp(q \cdot q^+)}{\exp(q \cdot q^+) + \sum_{q^-} \exp(q \cdot q^-)} \tag{1}$$

where $q \cdot q^+$ is the dot product between two vectors. Recently, CL has been introduced to various fields such as visual representation learning, visual question answering and image captioning with great success. In this paper, we introduce CL into a multi-subtask joint learning framework to learn cross-modal relationships through Visual-textual CL, and learn the relationships between answers and rationales through Answer-rationale CL to gain hidden commonsense semantic for VCR.

3 Proposed Approach

In VCR task, the dataset S can be formalized as a collection of n quadruples $\{V_i, Q_i, A_i, R_i\}_{i=1}^n$, where $V_i \in V$ is an image, $Q_i \in Q$ is a question about image V_i, $A_i \in A$ is the set of four candidate answers to question Q_i, and $R_i \in R$ is the set of four candidate rationales. Current works optimize two functions for question answering subtask and answer justification subtask respectively: (1) $f_{q2a} : V \times Q \rightarrow \mathbb{R}^{|A_i|}$ to produce correct answer and (2) $f_{qa2r} : V \times Q \times \tilde{A} \rightarrow \mathbb{R}^{|R_i|}$ to produce correct rationale, where $\tilde{A}$ is the correct answer set. Different from these works, we jointly train two subtasks with one model to produce correct answer and correct rationale simultaneously through learning the correlation between subtasks. Formally, we define the VCR task as follows:

$$\tilde{a_i} = \underset{a_{i,j} \in \mathcal{A}_i}{argmax} P(a_{i,j} | V_i, Q_i; \theta, \theta_a) \tag{2}$$

$$\tilde{r_i} = \underset{r_{i,k} \in \mathcal{R}_i}{argmax} P(r_{i,k} | V_i, Q_i; \theta, \theta_r) \tag{3}$$

where θ denotes the shared learnable parameters to both subtasks, and θ_a and θ_r are their respective parameters. Given image V_i and question Q_i, the model predicts the answer $\tilde{a_i} \in \mathcal{A}_i$ and the rationale $\tilde{r_i} \in \mathcal{R}_i$ simultaneously via jointly maximizing the two probabilities in Eq. (2) and Eq. (3).

We propose a Inter-subtask Consistent Representation Learning (ICRL) framework to jointly predict the answers and rationales for VCR. The architecture of ICRL is shown in Fig. 2 consisting of three parts: 1) Feature Extraction, 2) Multi-level Contrastive Learning, and 3) Classification.

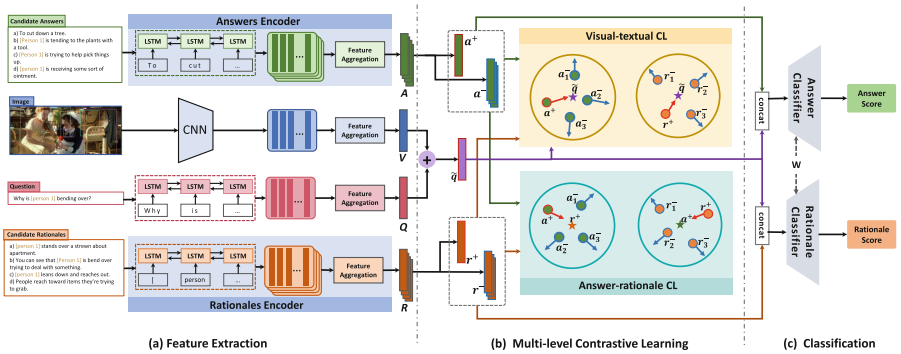

Fig. 2. The overall architecture of ICRL, which contains feature extraction module, multi-level contrastive learning module (visual-textual CL and answer-rationale CL), and classification module.

3.1 Joint Learning Framework

We design a joint learning framework to lay two subtasks in the same model structure to jointly optimize the model and obtain the prediction results of the two subtasks simultaneously. In the joint learning framework, the two streams of question answering and answer justification interact by sharing partial representations and parameters, as shown in Fig. 2.

In the feature extraction part, our model shares the representations of images and questions between subtasks. Furthermore, the textual encoders for candidate answers and candidate rationales share parameters. We then build a contrastive learning structure between answers and rationales for interaction between the two subtasks, which are described in more detail below. Moreover, the framework shares the parameters of the answer classifier and the rationale classifier for joint prediction in the classification module.

It should be pointed out that, unlike prior works which take the correct answer as an input for the answer justification subtask, the answer justification stream in our joint learning framework is not explicitly fed with correct answer, but obtain answer information from the interaction with question answering stream. In this way, there is a mutually reinforcing effect between answer prediction and rationale prediction.

3.2 Feature Extraction

In the feature extraction stage, we extract visual features for image and textual features for question, candidate answers and candidate rationales.

Given image I, we first obtain n visual region features $\{v_i\}_{i=1}^{n}$ from a pre-trained Faster R-CNN [14] with ResNet-101 [8] backbone. Thereafter, the visual feature is fused with its position feature p_i to get new visual region feature $\hat{v}_i \in \mathbb{R}^d$ as follows:

$$\hat{v}_i = \sigma(W_{vp}((W_v v_i + b_v)||(W_p p_i + b_p)) + b_{vp}) \tag{4}$$

where W_v, W_p and W_{vp} are weights, b_v, b_p and b_{vp} are biases, $\sigma(\cdot)$ denotes the activation function, and $(\cdot||\cdot)$ denotes the concatenation operation. Following [24], we then obtain a integrated image representation $V \in \mathbb{R}^d$ through a visual feature aggregation module:

$$V = \sum_{i=1}^{n} \alpha_i \hat{v}_i, \quad \alpha_i = \frac{\exp\left(\sigma\left(W_v^{'}\hat{v}_i + b_v^{'}\right)\right)}{\sum_{j=1}^{n} \exp\left(\sigma\left(W_v^{'}\hat{v}_j + b_v^{'}\right)\right)} \tag{5}$$

where $W_v^{'}$ is the weight, $b_v^{'}$ is the bias, $\sigma(\cdot)$ represents the activation function and $\alpha = [\alpha_1, \alpha_2, ..., \alpha_n] \in \mathbb{R}^n$ are the learned weights.

As for the textual modality, we adopt pre-trained BERT [6] to obtain context-enhanced embeddings from the given text, which are then fed into bidirectional LSTM to produce textual features $\{q_i\}_{i=1}^{l_q}$, $\{a_i\}_{i=1}^{l_a}$ and $\{r_i\}_{i=1}^{l_r}$, where $q_i \in \mathbb{R}^d$, $a_i \in \mathbb{R}^d$ and $r_i \in \mathbb{R}^d$ correspond to the i-th token in the question, each candidate answer and each candidate rationale respectively, and l_q, l_a and l_r denote sequence length. Note that the parameters of bidirectional LSTM for candidate answers and candidate rationales are all shared. Afterwards we use textual feature aggregation module to generate global textual representations $Q \in \mathbb{R}^d$, $A_k \in \mathbb{R}^d$, $R_k \in \mathbb{R}^d$ for each sentence through similar steps as in Eq. (5), where $k \in \{1, 2, 3, 4\}$ denotes four candidates.

3.3 Multi-level Contrastive Learning

The Multi-level Contrastive Learning network shown in Fig. 2 consists of two modules: Answer-rationale CL and Visual-textual CL.

Answer-Rationale CL. The Answer-rationale CL is designed to achieve semantic alignment between answers and rationales thus learn consistent representations for subtasks. First, we project answers and rationales into a common logical space and get their respective representations $\tilde{A} \in \mathbb{R}^{4\times d'}$ and $\tilde{R} \in \mathbb{R}^{4\times d'}$:

$$\tilde{A} = \{\tilde{A}_k\}_{k=1}^{4}, \quad \tilde{A}_k = \sigma(W_{la}A_k + b_{la}) \tag{6}$$

$$\tilde{R} = \{\tilde{R}_k\}_{k=1}^{4}, \quad \tilde{R}_k = \sigma(W_{lr}R_k + b_{lr}) \tag{7}$$

where $\tilde{A}_k \in \mathbb{R}^{d'}$ and $\tilde{R}_k \in \mathbb{R}^{d'}$ denotes the k-th candidate answer and rationale representation in the common logic space respectively, $W_{la} \in \mathbb{R}^{d'\times d}$ and $W_{lr} \in \mathbb{R}^{d'\times d}$ are two linear projection matrices, b_{la} and b_{kr} are biases, and $\sigma(\cdot)$ represents the activation function. We then separate out the correct answer $a^+ \in \mathbb{R}^{d'}$ and wrong answers $\{a_k^-\}_{k=1}^{3} \in \mathbb{R}^{3\times d'}$ from $\tilde{A}$. Similarly, the correct rationale $r^+ \in \mathbb{R}^{d'}$ and wrong rationales $\{r_k^-\}_{k=1}^{3} \in \mathbb{R}^{3\times d'}$ are separated from $\tilde{R}$. In the logical feature space, the representations of correct answer and correct rationale should be close, while the presentations of the correct answer and the wrong rationales, and the presentations of the correct rationale and the wrong answers should be much farther. Therefore, we treat a^+ as an anchor, (a^+, r^+) as

the positive pair and take $\{(a^+, r_1^-), (a^+, r_2^-), (a^+, r_3^-)\}$ as three negative pairs for contrastive learning. The contrastive loss is defined as follows:

$$L_{a2r} = -log \frac{\exp\left(s\left(a^+, r^+\right)/\tau\right)}{\exp\left(s\left(a^+, r^+\right)/\tau\right) + \sum_{k=1}^{3} \exp\left(s\left(a^+, r_k^-\right)/\tau\right)} \tag{8}$$

where $s(\cdot, \cdot)$ is dot product evaluating the similarity of two vectors, and τ is a hyperparameter controlling the sensitivity of $s(\cdot, \cdot)$. Through the similar step, when treat r^+ as anchor, we can obtain another contrastive loss:

$$L_{r2a} = -log \frac{\exp\left(s\left(r^+, a^+\right)/\tau\right)}{\exp\left(s\left(r^+, a^+\right)/\tau\right) + \sum_{k=1}^{3} \exp\left(s\left(r^+, a_k^-\right)/\tau\right)} \tag{9}$$

where the notations have the same meaning as in Eq. (8). Through the training, as shown in the yellow rectangle in Fig. 2, the features for matched answer-rationale pairs are directed to be close while the features for mismatched answer-rationale pairs are farther apart. Based on this, the model can ensure the semantic consistency of answers and rationales in feature space.

Visual-Textual CL. The visual-textual contrastive learning module is designed to model cross-modal relationships for VCR inspired by [24]. Specifically, we adopt element-wise adding operation followed by a linear mapping for fusing the visual representation V and question representation Q to get fused query representation $\tilde{q} \in \mathbb{R}^{d'}$. After that, similar to Answer-rationale CL, adopting $\tilde{q}$ as the anchor, correct options as positive samples and wrong options as negative samples, the contrastive losses of Visual-textual CL are formulated as follows:

$$L_{v2a} = -log \frac{\exp\left(s\left(\tilde{q}, a^+\right)/\tau\right)}{\exp\left(s\left(\tilde{q}, a^+\right)/\tau\right) + \sum_{k=1}^{3} \exp\left(s\left(\tilde{q}, a_k^-\right)/\tau\right)} \tag{10}$$

$$L_{v2r} = -log \frac{\exp\left(s\left(\tilde{q}, r^+\right)/\tau\right)}{\exp\left(s\left(\tilde{q}, r^+\right)/\tau\right) + \sum_{k=1}^{3} \exp\left(s\left(\tilde{q}, r_k^-\right)/\tau\right)} \tag{11}$$

where $s(\cdot, \cdot)$ and τ are as in Eq. (8). As shown in the green rectangle in Fig. 2, by minimizing these losses, the representation $\tilde{q}$ and the ground-truth responses representations are trained to be near in feature space, while the features with different semantics are trained to be farther. Therefore, Visual-textual CL helps the model achieve cross-modal semantic alignment.

3.4 Classification and Loss

We construct classifiers in the classification module to predict answers and rationales. For the question answering stream, we first use a concatenation operation

to fuse the representations of the query $\tilde{q}$ and the candidate answer A_k. We then feed the fused representation into a 1-layer MLP classifier and use a typical cross-entropy loss for classification to get answer prediction loss L_a. The answer justification stream is handled in the same way to get rationale prediction loss L_r. The final loss is defined as follows:

$$L = L_a + L_r + \lambda_1(L_{a2r} + L_{r2a}) + \lambda_2(L_{v2a} + L_{v2r}) \tag{12}$$

where λ_1 and λ_2 are the trade-off parameters. Jointly training the classification losses and the contrastive losses, our model can learn inter-subtask consistent representations enabling efficient reasoning for VCR. During test, given an image, a question, four candidate answers and four candidate rationales, we feed them into the trained ICRL network to get the prediction probabilities of answers and rationales and select the best answer and the best rationale as the final results.

4 Experiments

In the present section, we first present the VCR dataset [23] and the implementation details of the proposed method. We then compare the experimental results of ICRL with the state-of-the-arts. Finally, we perform ablation studies to explore the effectiveness of every module in ICRL.

4.1 Datasets and Implementation Details

We carry out experiments on VCR dataset [23] containing 290 K multiple-choice questions for both answers and rationales. Each image in the VCR dataset corresponds to several questions. Most of these questions are related to complex life scenarios and require commonsense-related reasoning to get answers. There are around 213 k questions in the training set, 27 k in the validation set and 1 5k in the test set of VCR. As mentioned above, we jointly train two subtasks (i.e., $Q \rightarrow A$ and $QA \rightarrow R$) in the same framework sharing the visual representations, question representations, and the parameters of the textual encoders. The prediction result of $Q \rightarrow AR$ is correct only if both $Q \rightarrow A$ and $QA \rightarrow R$ predict correctly. The activation functions in Eq. (4) is ReLU function. Equation (6) and Eq. (7) use LeakyReLU activation functions. The temperature τ in Eq. (8), Eq. (9), Eq. (10) and Eq. (11) is set 0.2, 0.2, 0.1 and 0.1, respectively. In Eq. (12), λ_1 is 0.1 and λ_2 is 0.7. We set the batch size to 96 and train the model for 25 epochs. The Adam optimizer [11] with 0.00005 weight decay and 0.9 beta are used to optimize our model. The initial learning rate is set to 0.0001, and multiplied by 0.2 when the validation accuracy does not increase. Pytorch is used to implement the proposed method and all ablated versions. We evaluate the model using accuracy as metric.

Table 1. Performance comparison between ICRL and the state-of-the-art baselines on VCR. **Best** results are highlighted in each column. (Since the test accuracies of VC-RCNN and CL-VCR are not reported in their work, and the test accuracy of VCR dataset can only be obtained by uploading the prediction results to the VCR leaderboard, we only show their validation accuracies.)

	Method	$Q \rightarrow A$		$QA \rightarrow R$		$Q \rightarrow AR$	
		Val	Test	Val	Test	Val	Test
Text-only	BERT	53.8	53.9	64.1	64.5	34.8	35.0
	BERT (response only)	27.6	27.7	26.3	26.2	7.6	7.3
	ESIM + ELMO	45.8	45.9	55.0	55.1	25.3	25.6
	LSTM + ELMO	28.1	28.3	28.7	28.5	8.3	8.4
VQA	RevisitedVQA	39.4	40.5	34.0	33.7	13.5	13.8
	BottomUpTopDown	42.8	44.1	25.1	25.1	10.7	11.0
	MLB	45.5	46.2	36.1	36.8	17.0	17.2
	MUTAN	44.4	45.5	32.0	32.2	14.6	14.6
VCR	R2C(2019)	63.8	65.1	67.2	67.3	43.1	44.0
	HGL(2019)	69.4	70.1	70.6	70.8	49.1	49.8
	CCN(2019)	67.4	68.5	70.6	70.5	47.7	48.4
	TAB-VCR(2019)	69.9	70.4	72.2	**71.7**	50.6	50.5
	CKRM(2020)	66.2	66.9	68.5	68.5	45.6	45.9
	VC-RCNN(2020)	67.4	–	69.5	–	–	–
	CL-VCR(2021)	69.9	–	70.6	–	–	–
	ICRL(Ours)	**71.0**	**70.7**	**72.7**	71.6	**51.9**	**50.9**

4.2 Performance Comparison

We compare the performance of ICRL with the following three types of methods to evaluate the effectiveness of the proposed model:

1) Text-only baselines in [23], including BERT [6], BERT (response only), ESIM+ELMo [4] and LSTM+ELMo. By comparing with this kind of methods, we can validate the effect of our model on visual content in representation learning.
2) VQA-based baselines, including Bottom-up and Top-down attention (BottomUpTopDown) [1], RevisitedVQA [9], Multimodal Low-rank Bilinear Attention (MLB) [10], and Multimodal Tucker Fusion (MUTAN) [2]. By comparing with such methods, we can verify the complex reasoning ability of our model.
3) Methods specially designed for the VCR, including R2C [23], CCN [19], TAB-VCR [13], CKRM [18], HGL [21], VC-RCNN [16] and CL-VCR [20]. The ability of our model to perform complex reasoning can be verified by comparison with these methods.

As with previous VCR-specific works, we do not compare ICRL with models pre-training on large external corpora such as UNITER [5] and SGEITL [17] since our approach does not rely on out-of-domain datasets. The results are shown in Table 1.

Our proposed approach obtains best performance compared with state-of-the-art methods on both validation set and test set of VCR dataset. On validation set, our approach gains accuracy of 71.0%, 72.7% and 51.9% for $Q \rightarrow A$, $QA \rightarrow R$ and $Q \rightarrow AR$ subtasks, respectively. Specifically, compared with the text-only methods, our approach obtains an improvement of 17.1% to 44.3% on the $Q \rightarrow AR$ subtask, which suggests the importance of visual content on the VCR task. Our model improves 34.9% to 41.2% on $Q \rightarrow AR$ over the VQA methods. This reflects the lack of inference ability of VQA methods for VCR task. Compared with VCR specific methods, our approach still achieves the best, which is 1.1% higher on $Q \rightarrow A$, 0.5% higher on $QA \rightarrow R$ and 1.3% higher on $Q \rightarrow AR$ than the second. Moreover, although R2C, HGL, CCN, and CKRM utilize sophisticated attention mechanisms, our proposed method still outperforms them. This is probably because our method can achieve not only cross-modal alignment but also learn inter-subtask consistent representations, which helps the model extract more reasonable and discriminative features for VCR.

4.3 Ablation Studies

To explore the effect of each module in our model, we conduct different ablation model.

- **Base**. The *Base* version handles the two subtasks of question answering and answer justification independently like previous works. The *Base (correct answer)* model takes the right answer as an input to the answer justification subtask. Because our model does not explicitly take the correct answer as an input for answer justification, we design the *base (predict answer)* version which takes predicted answer rather than the right answer as an input for fair comparison.
- **Base + Joint L**. *Base + Joint L* means joint learning of two subtasks based on the *Base* network. Note that the *Base + Joint L* architecture does not explicitly taken the correct answer or predicted answer as input.
- **Base + Joint L + Visual-textual CL**. This variant adds Answer-rationale CL into *Base + Joint L*, which can perform contrastive learning between image representation and option representations.
- **Base + Joint L + Answer-rationale CL**. These variant adds Answer-rationale CL into *Base + Joint L*, which can perform contrastive learning between answer representations and rationale representations.

Table 2. Ablation analysis of our proposed model over validation split. Base: only one subtask loss; Joint L: both subtasks losses; Visual-textual CL: visual-textual contrastive learning loss; Answer-rationale CL: answer-rationale contrastive learning loss.

Method	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
Base (correct answer)	69.11	72.01	50.07
Base (predict answer)	69.11	70.75	49.12
Base + Joint L	69.42	71.52	49.69
Base + Joint L + Visual-textual CL	70.65	72.40	51.27
Base + Joint L + Answer-rationale CL	69.71	71.41	49.99
ICRL	**71.04**	**72.70**	**51.86**

The results of ablation analysis of $Q \rightarrow A$, $QA \rightarrow R$ and $Q \rightarrow AR$ subtasks on VCR validation set are shown in Table 2. From the table, we conclude the following:

1) Compared with *Base (predict answer)*, *Base (correct answer)* has a slightly better effect on $QA \rightarrow R$ subtask, which reflects that the correct answer has an important positive effect on rationale prediction.
2) *Base + Joint L* has improved performance compared to *Base (predict answer)*. This implies a mutually reinforcing influence between the two subtasks, validating the effectiveness of joint learning for VCR.
3) The comparison between *Base + Joint L + Visual-textual CL* and *Base + Joint L* shows that visual-textual contrastive learning can help model learn discriminative representations by learning cross-modal semantics.
4) Compared with *Base + Joint L*, *Base + Joint L + Answer-rationale CL* also improves the effect, which verifies that Answer-rationale CL helps the model obtain better prediction results by learning inter-subtask consistent representations.
5) As compared to *Base + Joint L*, the performance of our overall model is significantly improved, which further validates the effectiveness of multi-level contrastive learning module.

5 Conclusion

In this paper, we proposed a novel inter-subtask consistent representation learning framework to learn inter-subtask hidden commonsense semantics for VCR. Specifically, the proposed multi-level contrastive learning module learned consistent representations by achieving cross-modal and inter-subtask semantic alignment, so as to help the model carry out effective reasoning for VCR. Extensive experiments on VCR dataset demonstrated the effectiveness of proposed approach. We plan to generalize our approach to more general inter-task learning in the future.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR, pp. 6077–6086 (2018)
2. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: multimodal tucker fusion for visual question answering. In: ICCV, pp. 2612–2620 (2017)
3. Bromley, J., et al.: Signature verification using a "siamese" time delay neural network. IJPRAI **7**(04), 669–688 (1993)
4. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., Inkpen, D.: Enhanced lstm for natural language inference. In: ACL, pp. 1657–1668 (2017)
5. Chen, Y.-C., et al.: UNITER: UNiversal image-TExt representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 104–120. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_7
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp. 4171–4186 (2018)
7. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR, vol. 2, pp. 1735–1742 (2006)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
9. Jabri, A., Joulin, A., van der Maaten, L.: Revisiting visual question answering baselines. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 727–739. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_44
10. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. In: ICLR (2017)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
12. Lee, J., Kim, I.: Vision-language-knowledge co-embedding for visual commonsense reasoning. Sensors **21**(9), 2911 (2021)
13. Lin, J., Jain, U., Schwing, A.: Tab-vcr: tags and attributes based vcr baselines. In: NIPS, pp. 15589–15602 (2019)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. TPAMI **39**(6), 1137–1149 (2016)
15. Song, D., Ma, S., Sun, Z., Yang, S., Liao, L.: Kvl-bert: knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. Knowl.-Based Syst. **230**, 107408 (2021)
16. Wang, T., Huang, J., Zhang, H., Sun, Q.: Visual commonsense r-cnn. In: CVPR, pp. 10760–10770 (2020)
17. Wang, Z., et al.: Sgeitl: scene graph enhanced image-text learning for visual commonsense reasoning. arXiv (2021)
18. Wen, Z., Peng, Y.: Multi-level knowledge injecting for visual commonsense reasoning. TCSVT **31**(3), 1042–1054 (2020)
19. Wu, A., Zhu, L., Han, Y., Yang, Y.: Connective cognition network for directional visual commonsense reasoning. In: NIPS, pp. 5669–5679 (2019)
20. Ye, K., Kovashka, A.: A case study of the shortcut effects in visual commonsense reasoning. In: AAAI, pp. 3181–3189 (2021)
21. Yu, W., Zhou, J., Yu, W., Liang, X., Xiao, N.: Heterogeneous graph learning for visual commonsense reasoning. In: NIPS, pp. 2765–2775 (2019)

22. Zamir, A.R., et al.: Robust learning through cross-task consistency. In: CVPR, pp. 11197–11206 (2020)
23. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: visual commonsense reasoning. In: CVPR, pp. 6720–6731 (2019)
24. Zhang, X., Zhang, F., Xu, C.: Multi-level counterfactual contrast for visual commonsense reasoning. In: MM, pp. 1793–1802 (2021)