DOI: 10.19884/j.1672-5220.202403004

Context-Aware Visual Entailment Driven by Specific Instructions

HAN Yufeng[1,2], HAO Kuangrong[1,2*], TANG Xuesong[1,2], WEI Bing[1,2]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile and Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: Visual entailment (VE) is a prototypical task in multimodal visual reasoning, where current methods frequently utilize large language models (LLMs) as the knowledge base to assist in answering questions. These methods heavily rely on the textual modality, which inherently cannot capture the full extent of information contained within images. We propose a context-aware visual entailment (CAVE) model, which introduces a novel aggregation module designed to extract high-level semantic features from images. This module integrates lower-level semantic image features into high-level visual tokens, formatting them similarly to text tokens so that they can serve as inputs for LLMs. The CAVE model compensates for the loss of image information and integrates it more effectively with textual comprehension. Additionally, the CAVE model incorporates a new input format and training methodology, which is rooted in instruction tuning and in-context learning techniques. The objective of this research is to maximize the inherent logical reasoning capabilities of LLMs. Experimental results on the E-SNLI-VE dataset show that the proposed CAVE model exhibits outstanding performance.

Keywords: visual entailment (VE); textual-visual integration; instruction tuning; in-context learning
CLC number: TP181 **Document code**: A
Article ID: 1672-5220(2025)02-0177-10

Open Science Identity (OSID)

0 Introduction

Multimodal visual reasoning[1-6] refers to a method that leverages multimodal information for answering visual questions. It aims to understand and reason about images by jointly encoding and learning the associations between visual and linguistic cues. Currently, there are three paradigms of multimodal visual reasoning models.

1) Explainer-based explicit models[7-10]. These models utilize convolutional neural networks (CNNs)[11] to extract prominent image features and incorporate attention mechanisms to prioritize the image features that are pertinent to the question. They explicitly model the logical relationship between the image and the text question, and predict reasons using an explainer such as long short-term memory (LSTM)[12] or generative pre-trained transformer 2 (GPT-2)[13].

2) Two-stage pre-trained transformer models[2,14-17]. These models consist of a pre-training stage followed by fine-tuning. Initially, they learn the joint semantic features of images and texts through contrastive learning on a large-scale and noisy dataset of image-text pairs. Then, they undergo fine-tuning using high-quality annotated data from specific downstream tasks. These models significantly improve performance and generalizability but remain close-set and domain-dependent, with limited answer diversity and a tendency to generate only short texts linked to predefined visual-textual tasks.

3) Fine-tuning large-scale language models[3,18-21]. These models utilize powerful pre-trained language models (such as GPT-3[22]) to adjust to different scenarios and needs in multimodal applications. Gui et al.[18] and Lin et al.[19] proposed novel architectures that utilized GPT-3 as a knowledge engine to assist in multimodal visual reasoning, even though they functioned solely in the domain of text-based question answering. However, relying only on captions without direct image analyses can lead to the model erroneously determining the entailment relationship between texts and images. Insufficient or misleading information within captions might cause the model to inappropriately classify a text as not entailing or irrelevant to an image. For example, this occurs when a crucial detail in the image is left out of the caption but is mentioned in the text. Consequently, the model may make judgments based solely on caption data, potentially misinterpreting the true visual-linguistic alignment.

To address this issue, we propose an aggregation module that extracts high-level semantic features from images as visual tokens, aligning the image and text domains. This implicit approach aims to alleviate the challenges posed by visual ambiguities and missing

Received date: 2024-03-12
Foundation items: Fundamental Research Funds for the Central Universities, China (No. 2232021A-10); Shanghai Pujiang Program, China (No. 22PJ1423400)
* Correspondence should be addressed to HAO Kuangrong, email: krhao@dhu.edu.cn
Citation: HAN Y F, HAO K R, TANG X S, et al. Context-aware visual entailment driven by specific instructions[J]. *Journal of Donghua University (English Edition)*, 2025, 42(2): 177-186.

semantic information, ensuring a more comprehensive understanding of the interplay between visual content and textual description. We enhance the model's logical reasoning ability by utilizing specific instructions and in-context learning techniques to leverage the text reasoning power of large language models (LLMs) in multimodal tasks. The contrastive language-image pre-training (CLIP)[23] and the aggregation layer jointly operate as an image encoder, effectively transforming images into visual tokens. Vicuna[24] is used as a text decoder, with inputs that include visual tokens, image captions, hypotheses and answers. These elements are meticulously adjusted according to specific instructions to construct coherent and comprehensive sentences. These sentences are then input into Vicuna to generate the predicted answer.

Overall, the contributions are as follows.

1) A novel context-aware visual entailment (CAVE) model is presented, which integrates CLIP and Vicuna, and this model converts a classification task into a generation task.

2) An aggregation module is proposed, which extracts high-level semantic features from images, converting them into virtual tokens, thereby aligning the image and text modules.

3) An efficient instruction is designed for input samples, and in-context learning is utilized during the training process to optimize and enhance model performance.

4) The proposed model has been experimentally validated and its substantial performance on the E-SNLI-VE dataset[25] has been demonstrated.

1 Related Works

1.1 Explainer-based explicit models

Figure 1 provides an illustrative overview of the proposed CAVE model.

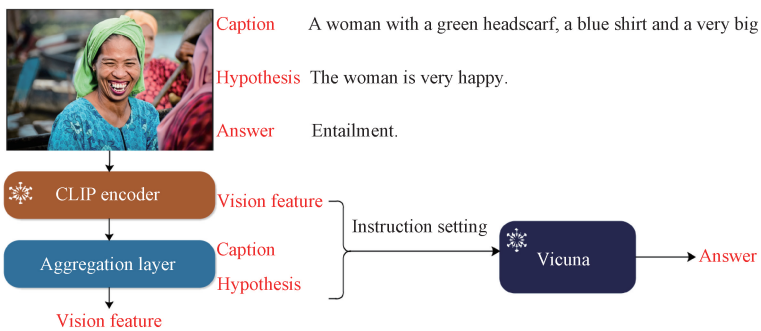

Fig. 1 An illustrative overview of the proposed CAVE model

Researchers encoded images and questions to obtain joint features, learned attentional characteristics from questions to images, and used LSTM to predict reasons[7-8]. Thomas et al.[9] employed multi-instance learning to make fine-grained predictions of the logical relationships between knowledge elements in images and texts. Although this method demonstrates the modeling of semantic relationships between images and texts, the explainable generator falls short in its ability to comprehend the global context and interact with it effectively, resulting in suboptimal integration of cross-modal information.

1.2 Two-stage pre-trained transformer models

Researchers have made significant advancements in recent years within the field of artificial intelligence, particularly at the intersection of natural language processing (NLP) and computer vision (CV)[14-15, 17]. These models typically consist of two main stages: pre-training and fine-tuning. In the pre-training stage, the model is trained using large-scale unlabeled or weakly-labeled pairs of images and texts, often sourced from the Internet or other extensive datasets where there may be some association between the images and texts, but not always precise alignment. The primary goal during this stage is to enable the model to learn joint semantic features between images and texts. This is commonly achieved through contrastive learning, where the model learns to distinguish between matching image-text pairs and non-matching pairs. To achieve this, the model creates an embedding space where corresponding image and text representations are brought closer together, while disparate ones are pushed apart. During the fine-tuning stage, the model is trained using high-quality annotated data for downstream tasks. These tasks usually relate to specific application scenarios such as image captioning, visual question answering (VQA), and image retrieval. At this stage, the model adjusts its parameters based on the task requirements to optimize performance on that particular dataset. While these models have shown improvements in both performance and generalization, they still have certain limitations. Notably, neural networks tend to be closed-set and domain-dependent. It means that they are good at new tasks which are similar to those encountered during training. However, their performance may deteriorate when faced with tasks that are substantially different from the distribution of the training data.

1.3 Fine-tuning large-scale language models

In the realm of multi-modal visual reasoning, several methods leveraging LLMs have emerged, such as in VQA tasks where Yang et al.[26] used captions generated from images and questions directly inputted into the LLM to produce answers. Additionally, Gui et al.[18] and Lin et al.[19] utilized the knowledge about questions

and captions within the LLM to serve as candidate justifications for assisting in the training of VQA models. However, captions provide limited descriptions that fail to capture the full informational content of an image. Therefore, we propose a novel aggregation module to extract higher-level semantic features from images, compensating for the information gap in captions. Simultaneously, we explore a new input scheme and training strategy based on specific instructions and in-context learning to fully leverage the understanding and reasoning capabilities of LLMs in completing visual entailment (VE) tasks.

2 Methods

2.1 Problem description

Given an image I as the premise, a simple caption of the image c, and a sentence h as the hypothesis, the goal is to determine whether the image semantically entails, neutral concerning or contradicts the hypothesis. There are three possible outcomes a: entailment, neutrality and contradiction. Consequently, each instance in the dataset can be denoted as a tuple $\{I, c, h, a\}$.

2.2 Model architecture

The proposed CAVE model consists of three key components: an aggregation-based image encoder which includes a CLIP encoder and an aggregation layer; a context-aware text decoder consisting of Vicuna and several contextual samples; an instruction-guided cross-modal information integration operation. The overall architecture of the proposed CAVE model is illustrated in Fig. 2, where I_1^i, I_2^i, $\cdots$, I_s^i $(i=1, 2, \cdots, N)$ denote the ith given image patch features, and I_1^t, I_2^t, $\cdots$, I_s^t denote the image patch features for inference.

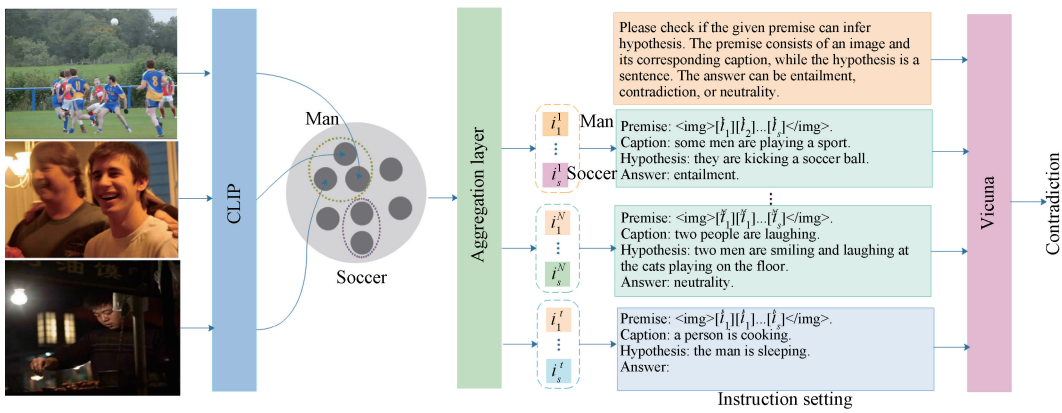

Fig. 2 Overall architecture of the proposed CAVE model

2.2.1 Aggregation-based image encoding

The model is built upon the foundation of CLIP's image encoder and an innovative aggregation network. CLIP, introduced by Radford et al.[23] in 2021, is a self-supervised learning framework that excels particularly in image classification tasks. This model integrates an image encoder and a text decoder. Through large-scale training on aligned image-text pairs, it achieves instance-level cross-modal feature matching, demonstrating remarkable generalization capabilities. Within this module, we utilize the CLIP image encoder f_{CLIP} to extract visual features $\boldsymbol{v}$, and mathematically express its application as

$$\boldsymbol{v} = f_{\text{CLIP}}(I), \tag{1}$$

where $\boldsymbol{v} \in \mathbf{R}^{N_1 \times D}$ represents the un-pooled features from the last layer of the CLIP image encoder that is flattened out; N_1 and D represent the dimensions of $\boldsymbol{v}$. These patch-level features provide rich semantic information as input to the downstream aggregation project module.

VE tasks demand a more comprehensive set of image features to provide the necessary context for determining whether a hypothesis holds, necessitating further refinement of high-level semantic visual features and effective mapping of these into the textual feature space. We design an innovative aggregation mapping architecture, as depicted in Fig. 3, which incorporates two complementary pathways. On one pathway, a linear transformation layer FC $(\cdot)$, followed by a softmax function Softmax $(\cdot)$ and a transpose operation, is employed to perform the process of feature selection and aggregation, thereby yielding a feature selection matrix $\boldsymbol{m} \in \mathbf{R}^{N_2 \times N_1}$, where N_1 and N_2 represent the dimensions of $\boldsymbol{m}$.

$$\boldsymbol{m} = \text{Softmax}(\text{FC}(\boldsymbol{v}))^{\text{T}}. \tag{2}$$

This matrix effectively condenses the essential information from the image into a structured representation that can be more closely aligned with the textual modality. On the other pathway, the multi-layer perceptron is used to support the deep transformation of features, enhancing their expressive capacity and generating candidate features $\boldsymbol{v}_{\text{can}} \in \mathbf{R}^{N_1 \times D}$.

$$\boldsymbol{v}_{\text{can}} = \text{MLP}(\boldsymbol{v}), \tag{3}$$

where MLP refers to a function that transforms the input features $\boldsymbol{v}$ into more expressive features $\boldsymbol{v}_{\text{can}}$.

This component enhances the feature selection process by incorporating a non-linear mapping that can capture more intricate relationships within the visual data.

The multi-layer perceptron essentially refines these image features into semantically richer representations, ensuring that they carry sufficient information for accurate entailment judgment when integrated with textual information in the context of VE tasks. Ultimately, the outputs from both pathways are combined to produce a novel aggregated feature representation $\boldsymbol{v}_a \in \mathbf{R}^{N_2 \times D}$.

$$\boldsymbol{v}_a = \boldsymbol{m} \cdot \boldsymbol{v}_{\text{can}}. \tag{4}$$

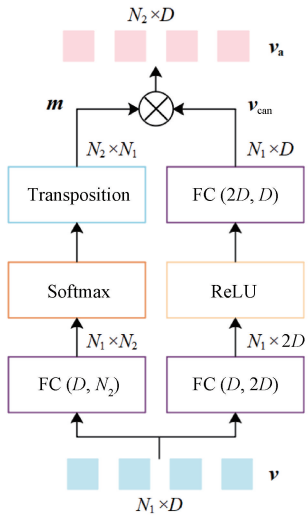

Fig. 3 Illustration of aggregation mapping module

2.2.2 Context-aware text decoding

With an extremely large number of parameters, the Vicuna model boasts an enhanced capacity for representation learning when tackling complex text understanding and generation tasks. The larger model capacity enables it to capture a broader range of linguistic rules, patterns and subtle semantic distinctions, thereby increasing the likelihood of generating high-quality text output. During its construction, the Vicuna model leverages vast amounts of diverse and heterogeneous textual data for pre-training, covering a wide array of topics and domains. This ensures that the model develops a profound understanding and broad adaptability to various types of textual content. Vicuna model is favored for its superior in-context learning ability and task adaptiveness. It can learn from just a few examples of input-output pairs without requiring fine-tuning when transferred to new tasks. The key to the LLM's performance lies in its attention layers, which facilitate an implicit parameter optimization process during inference, similar to explicit optimization through gradient descent during fine-tuning. Building upon this, the proposed model introduces in-context learning in two stages.

In the first stage, the aggregation network learns to infer a latent concept by utilizing the four components of a prompt: input, output, format and input-output mapping. These latent concepts encompass lexical distributions, formats and syntactic relationships, which means premise →hypothesis→relationship. In the second stage, despite examples being concatenated in a non-continuous manner, the LLM can still make predictions by leveraging shared concepts (premise→hypothesis→relationship), indicating that in-context learning has occurred.

Let the Vicuna model be denoted as f_{dec}, and then the output is depicted as

$$\boldsymbol{y} = f_{\text{dec}}(\text{format}(\boldsymbol{v}_a, \boldsymbol{x}_{\text{t1}}, \boldsymbol{x}_{\text{t2}}, \boldsymbol{x}_{\text{t3}})), \tag{5}$$

where format(·) represents the input setting, the first stage involves tensor concatenation, and the second stage encompasses the instruction step; $\boldsymbol{x}_{\text{t1}}$, $\boldsymbol{x}_{\text{t2}}$ and $\boldsymbol{x}_{\text{t3}}$ denote textual embedding vectors for captions, hypotheses and answers, respectively; $\boldsymbol{y} \in \mathbf{R}^{l_2 \times D_{\text{v}}}$ represents the prediction by the Vicuna language model, l_2 denotes the maximum length of the output tokens, and D_{v} refers to the size of the vocabulary.

2.2.3 Instruction-guided cross-modal information integration

In the cross-modal fusion phase, we incorporate principles of instruction learning by using a carefully designed set of instructions to guide Vicuna on how to dynamically combine features from both modalities based on task-specific requirements, as depicted in Fig. 4.

Here, [img1][img2]⋯[img$N$] represent visual tokens extracted from image features, while X_{t1}, X_{t2} and X_{t3} denote textual tokens for captions, hypotheses and answers, respectively. The black text represents fixed prompt information, whereas green-marked symbols such as <STOP> and <IMG>, and double-hash symbol (##) serve as auxiliary indicators that signal the start and end of dialogues, the boundaries of image markers and the conclusion of answer sections. These instructions not only specify how the image and text features interact but also define the emphasis on one modality's information or how to harmonize the relationship between them in specific application scenarios. For example, when an instruction requires the model to determine entailment based on a text description, the system will prioritize guiding text features while also integrating image features in its assessment. During an image classification task, the model may rely more heavily on image features while still referring to text descriptions to improve accuracy. Operationally, Vicuna incorporates specific instructions as additional inputs. It then adjusts weight assignments within self-attention layers or cross-modal attention layers using these instruction signals to achieve a dynamic fusion of the visual and textual feature vectors.

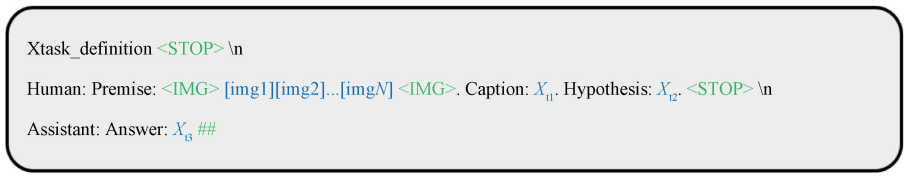

Fig. 4 Diagram of instruction step

2.2.4 Objective function

The objective is to minimize the discrepancy between the predicted probability distribution and the true probability distribution. Concretely, the proposed model predicts a probability distribution $\boldsymbol{p}_{ij} = (\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}, \cdots, \boldsymbol{p}_{il})$ for each sample $\boldsymbol{p}_i$, where $\boldsymbol{p}_{ij} \in \mathbf{R}^{D_v}$ denotes the prediction of the jth word. $\hat{\boldsymbol{y}}_i = (\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2}, \cdots, \boldsymbol{a}_{il})$ represents the label for the ith sample, where $\boldsymbol{a}_{ij}$ denotes the one-hot vector of the jth word. N denotes the number of the dataset; l denotes the maximum length for each sample; $i \in 1, 2, \cdots, N$; $j \in 1, 2, \cdots, l$. The loss function L is defined as

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l}\hat{\boldsymbol{y}}_j \log \boldsymbol{p}_{ij}. \tag{6}$$

2.3 Training details

We utilize original image inputs of size 224×224×3 and extract high-level visual features using the ViT-Base model from CLIP. Concurrently, a seven billion-parameter Vicuna-7B language model is utilized for text comprehension and generation. The training process is executed on an NVIDIA TITAN RTX GPU with a batch size set to 4 to strike a balance between computational resources and memory efficiency. For optimizing model performance, we utilize the AdamW optimizer[27] as the update strategy, initializing the learning rate at 3×10^{-5}. This value has been empirically determined through preliminary experimentation as the optimal starting point. Our training process consists of two periods. The initial training period allows the model to train the aggregation layer without instructions. During this period, we record and monitor performance metrics of the validation set at each iteration to identify the optimal model weights during the training process. After the initial training period, we enter the in-context learning period, during which we load the parameters from the optimal model and then freeze them for in-context learning. This enables the model to leverage its foundational knowledge to comprehend and adjust to contextual information unique to the task environment, thereby improving performance on VE tasks. Throughout the entire training regimen, techniques such as early stopping are also implemented to prevent overfitting and ensure the model's generalization capabilities.

3 Experimental Section

3.1 Dataset and metrics

E-SNLI-VE[25], serving as a cutting-edge benchmark dataset for the VE task, not only extends and refines SNLI-VE[28] by integrating image descriptions from Flickr30k[29] with VE principles but also, after rectifying annotation errors in the original dataset[30], provides 401 700 training samples, 14 300 validation samples and 14 700 test samples to gauge models' cross-modal understanding capabilities between images and texts.

Three advanced automated evaluation metrics are employed to evaluate model performance: metric for evaluation of translation with explicit ordering (METEOR)[31], bidirectional encoder representations from Transformers score (BERTScore)[32] and accuracy (Acc). METEOR provides a holistic assessment of translation quality by considering word matches, phrase matches and semantic similarity. Additionally, it quantifies the degree of lexical overlap between the generated and reference texts. BERTScore, which leverages the pre-trained language model BERT, computes semantic similarity between sentences and is widely used in assessing outcomes in areas such as text generation and machine translations. Acc assesses the classification performance of the proposed model. In the VE task, there are three prediction classes: entailment, neutrality and contradiction.

3.2 Baseline models

We compared our proposed model with several advanced baseline models, including pointing and justification explanation (PJ-X)[8], faithful multimodal explanation (FME)[7], Rationale^{VT} Transformer (RVT)[33], question-answering-only (QA-only)[25], e-UG[25], natural language explanations in vision and vision-language tasks (NLX-GPT)[34] and one for all (OFA-X)[16]. The performance of these models is evaluated in the E-SNLI-VE test dataset.

3.3 Quantitative analysis

Table 1 illustrates that CAVE achieves the highest scores across all evaluation metrics, indicating its ability in generating high-quality text that is semantically coherent and accurate. CAVE encompasses two variations: CAVE_ENC, where labels remain as the original text, entailment, neutrality and contradiction; CAVE_YNU, which modifies these labels to yes, unknown and no. CAVE_YNU outperforms CAVE_ENC by 1.5% in METEOR (M), 4.3% in BERTScore (BS), and 3.6% in Acc, respectively. Note that all percentage increases mentioned in this section are absolute improvements. The rationale behind this differentiation is that the original label format poses a greater challenge for generating the corresponding tokens in Vicuna. Words like "entailment" and "contradiction" are split into multiple tokens ("entail" and "ment", "contradict" and "tion", respectively). Assuming that each token predicts

the probability p_n, then the correct prediction of entailment or contradiction requires $p_1 \times p_2$, where p_1 is the probability of the first token and p_2 is the probability of the second token. Since CAVE_YNU only needs to predict the probability of one token (yes, unknown or no), it is simpler for the model.

Table 1 Comparative evaluation of model performance in E-SNLI-VE test dataset

Model	Metric score/%		
	M	BS	Acc
PJ-X[8]	14.7	79.1	69.2
FME[7]	15.6	79.7	73.7
RVT[33]	18.8	81.1	72.0
QA-only[25]	18.7	81.1	—
e-UG[25]	19.6	81.7	79.5
NLX-GPT[34]	18.8	80.8	73.9
OFA-X[16]	18.6	85.7	80.9
CAVE_ENC	19.0	86.3	83.2
CAVE_YNU	**20.5**	**90.6**	**86.8**

To further investigate the impact of the proposed individual components on the overall performance of the model, we conducted an ablation analysis, as illustrated in Table 2. The first row of data represents the proposed model as a baseline reference. Subsequently, one of the following components was removed (expressed as w/o in Table 2): image tokens, image captions, instructions or context learning (ICL) strategy. In the experiment involving the removal of image tokens, we trained the image features by replacing them with random vectors sampled from a normal distribution. It can be observed that all three metrics decrease, highlighting the significance of image tokens in model performance. After removing image captions, the accuracy drops significantly by 10.5%, while BERTScore and METEOR decrease by 5.8% and 2.8%, respectively. This shows that image captions are indispensable for modeling logical reasoning. After removing the ICL strategy, all three metrics decrease, but the decrease is not significant, and the impact on model performance is negligible. To summarize, image tokens and image captions contribute the most to the model's performance. The ICL strategy optimizes the model performance, demonstrating the effectiveness of our input sample settings.

Table 2 Ablation study

Model setting	Metric score/%		
	M	BS	Acc
CAVE_YNU	**20.5**	**90.6**	**86.8**
w/o image token	18.4	87.4	83.8
w/o image caption	17.7	84.8	76.3
w/o instruction	20.3	89.9	86.1
w/o ICL	20.4	90.1	85.9

Table 3 shows a comparison of the results of the mapping layer between image features and text features using different structures, including the linear layer, the self-attention layer, and the aggregation layer we designed. It can be seen that the aggregation layer achieves the highest accuracy, followed by the self-attention layer, with the linear layer performing the least effectively in terms of accuracy. In addition, the number of training parameters in the aggregation layer is lower than that in the self-attention layer, but higher than that in the linear layer. This proves that it is effective to introduce more complex nonlinear mapping when converting the original feature distribution of the image encoder and text decoder. The aggregation layer is superior to the self-attention layer, indicating that not all image candidate features are beneficial to downstream text feature alignment and inference. Our proposed aggregation layer is effective in this respect, and it can better fusion image information to extract high-level semantic features.

Table 3 Comparison results of mapping layer

Mapping layer	Training parameters	Metric score/%		
		M	BS	Acc
Linear	$\mathbf{3.8\times10^7}$	19.1	87.3	83.5
Self-attention	1.17×10^8	20.3	89.7	86.2
Aggregation (ours)	6.7×10^7	**20.5**	**90.6**	**86.8**

Table 4 shows the comparative experimental results of the CAVE model with different aggregate token lengths under two label settings: CAVE_ENC and CAVE_YNU. The results indicate that the CAVE_YNU configuration outperforms CAVE_ENC in terms of overall performance metrics. Furthermore, an intriguing trend is observed in relation to the token length: there is a notable decrease in performance metrics as the length of input tokens grows, which is more pronounced than what was initially expected. This suggests that longer inputs may introduce additional complexity or noise, adversely affecting model performance.

Table 4 Comparative experiments of different aggregate token lengths

Model	Token length	Metric score/%		
		M	BS	Acc
CAVE_ENC	5	19.0	86.3	83.2
CAVE_YNU		20.5	90.6	86.8
CAVE_ENC	10	18.9	85.6	82.0
CAVE_YNU		20.1	89.5	84.7
CAVE_ENC	15	18.8	84.5	80.2
CAVE_YNU		20.1	87.1	83.9

Figure 5 shows the performance comparison across different training epochs for ICL samples. It can be seen that when the number of training epochs equals three, the model performance reaches its peak.

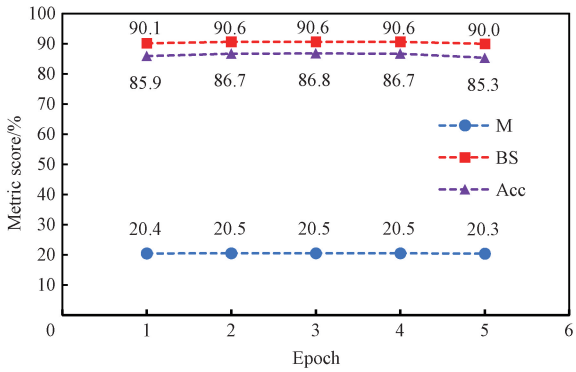

Fig. 5 Performance comparison for different ICL samples

3.4 Qualitative analysis

To visually show the results of converting image features into visual tokens, we select the five words that are most similar to visual tokens by calculating the cosine similarity between visual token vectors and all token vectors in the vocabulary. Then, these words are arranged in descending order of the similarity score, and visualized in the form of a word cloud map, as shown in Fig. 6. In the four pairs of images shown, the word with the highest similarity in each image is the main semantic information in the image, such as "sitting" "woman" and "young child". For example, in the first word cloud, we observe words such as "legs" "brown" "glass" and "hair", which describe the various object concepts in the image. It is worth noting that these concepts are in the graph. The description or hypothesis of the image is not mentioned in the text, thus highlighting the effectiveness of the aggregation network for capturing image semantic features. In the second image, we see words related to "button" "holding" "taking", etc., which further complement semantic information not fully covered in the caption description of the image. Overall, the visualizations in Fig. 6 once again validate that the aggregation network extracts effective and high-level semantic features.

Fig. 6 Word cloud visualization of visual tokens

In Fig. 7, we visualize the predicted answers. As can be seen, the first sentence of the predicted answer follows a template construction consistent with the template structure we provided in the training data. The LLM Vicuna plays a key role in this process. It enables our model to go beyond the traditional three-category classification task, establish the connection between the image and the text based on the image

features provided by the aggregation layer. Furthermore, it explains the reason based on the caption, demonstrating its ability in semantic understanding and logical reasoning.

	Hypothesis	The woman is very happy.
	Caption	A woman with a green headscarf, a blue shirt and a very big grin.
	Ground truth	Entailment.
	Prediction	Answer: entailment. The premise describe a specific with details of a woman is laughing.
	Hypothesis	A tall human jumping.
	Caption	A bearded man jumps in the snow with the ocean in the background and an orange, recumbent bike parked near a road sign.
	Ground truth	Entailment.
	Prediction	Answer: entailment. The premise describes a specific scenario involving a bearded man jumping.
	Hypothesis	A person is waiting for a train.
	Caption	A person in orange clothing rests above a metro entrance.
	Ground truth	Neutrality.
	Prediction	Answer: neutral. The premise simply describes a person in orange clothing above a metro entrance, without explicitly stating that the person is waiting for a train.
	Hypothesis	A man getting ready to play golf.
	Caption	A man in a black tank top wearing a red plaid hat.
	Ground truth	Neutrality.
	Prediction	Answer: neutral. The premise does not provide any direct information or implication about the man's activities or intentions related to playing golf.
	Hypothesis	The lioness is sleeping.
	Caption	A lioness closes in on its prey.
	Ground truth	Contradiction.
	Prediction	Answer: contradiction. The premise indicates that the lioness is actively hunting its prey, which contradicts the hypothesis that the lioness is sleeping.
	Hypothesis	The car is yellow.
	Caption	A biker races.
	Ground truth	Contradiction.
	Prediction	Answer: contradiction. The two statements are not logically connected, and both of them do not necessarily imply the hypothesis.

Fig. 7 Some qualitative examples from our model on the E-SNLI-VE test split

3.5 Strengths and limitations

3.5.1 Strengths

The main strengths of the proposed CAVE model are: the exploitation of cutting-edge pre-trained models (CLIP and Vicuna) to bolster cross-modal comprehension; the utilization of a tailored aggregation mechanism that facilitates smooth interaction and fusion of information across modalities; task specificity through instruction design, which guides and focuses the model's reasoning process for improved performance in targeted tasks.

3.5.2 Limitations

Despite these strengths, the proposed CAVE model is inherently constrained by its task-specificity, being primarily designed for logical reasoning in multimodal contexts. It does not readily generalize to other question-answering (QA) tasks without additional adaptation. More specifically, to broaden the model's applicability spectrum, there is a need to train a supplementary aggregation module capable of flexibly and effectively integrating various types of input data in different scenarios. This highlights the current limitation in terms of direct transferability to diverse QA tasks and represents a potential area for future improvement in the model's versatility and adaptability.

4 Conclusions

In this paper, we propose a novel model named CAVE, which integrates an established CLIP image encoder with an LLM, Vicuna, serving as the decoder. To facilitate the integration, we introduce a streamlined aggregation network that acts as a transformative interface between the visual and textual modalities. It not only extracts high-level semantic features of images but also converts them into visual tokens to align texts effectively. Also, the proposed model combines instruction tuning and in-context learning in the training process to improve the training efficiency. The qualitative experiments show our significant results on the E-SNLI-VE dataset. Nevertheless, multimodal visual reasoning still faces many challenges, such as math reasoning capabilities, which we will explore further in the future.

References

[1] DONG Q X, QIN Z W, XIA H M, et al. Premise-based multimodal reasoning: conditional

inference on joint textual and visual clues[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 932-946.
[2] CHEN Y C, LI L J, YU L C, et al. UNITER: universal image-text representation learning [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 104-120.
[3] SHAO Z W, YU Z, WANG M, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 14974-14983.
[4] SUZUKI R, YANAKA H, YOSHIKAWA M, et al. Multimodal logical inference system for visual-textual entailment[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Stroudsburg: ACL, 2019: 386-392.
[5] YU S J, JIN X Q, WU G W, et al. Deep multi-module based language priors mitigation model for visual question answering [J]. *Journal of Donghua University (English Edition)*, 2023, 40(6): 684-694.
[6] LI Q, TAO Q Y, JOTY S, et al. VQA-E: explaining, elaborating, and enhancing your answers for visual questions[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 570-586.
[7] WU J L, MOONEY R. Faithful multimodal explanation for visual question answering[C]// Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Stroudsburg: ACL, 2019: 103-112.
[8] PARK D H, HENDRICKS L A, AKATA Z, et al. Multimodal explanations: justifying decisions and pointing to the evidence[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 8779-8788.
[9] THOMAS C, ZHANG Y P, CHANG S F. Fine-grained visual entailment[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 398-416.
[10] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6077-6086.
[11] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. *Neural Computation*, 1989, 1(4): 541-551.
[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.
[13] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. (2019-02-14) [2024-03-15]. https://openai.com/blog/better-language-models/.
[14] LI J N, SELVARAJU R R, GOTMARE A D, et al. Align before fuse: vision and language representation learning with momentum distillation[EB/OL]. (2021-10-07) [2024-03-15]. https://arxiv.org/abs/2107.07651.
[15] ZENG Y, ZHANG X S, LI H. Multi-grained vision language pre-training: aligning texts with visual concepts[EB/OL]. (2022-06-01) [2024-03-15]. https://arxiv.org/abs/2111.08276.
[16] PLÜSTER B, AMBSDORF J, BRAACH L, et al. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations[EB/OL]. (2023-03-29) [2024-03-15]. https://arxiv.org/abs/2212.04231.
[17] WANG P, YANG A, MEN R, et al. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[EB/OL]. (2022-06-01) [2024-03-15]. https://arxiv.org/abs/2202.03052.
[18] GUI L K, WANG B R, HUANG Q Y, et al. KAT: a knowledge augmented transformer for vision-and-language [C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2022: 956-968.
[19] LIN Y Z, XIE Y J, CHEN D D, et al. REVIVE: regional visual representation matters in knowledge-based visual question answering [EB/OL]. (2022-10-10) [2024-03-18]. https://arxiv.org/abs/2206.01201.
[20] YANG H, LIN J Y, YANG A, et al. Prompt tuning for generative multimodal pretrained models[EB/OL]. (2022-08-04) [2024-03-15]. https://arxiv.org/abs/2208.02532.
[21] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning[EB/OL]. (2023-04-17) [2024-03-18]. https://arxiv.org/abs/2304.08485.
[22] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). New York: ACM, 2020: 1877-1901.
[23] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. Berlin: PMLR, 2021: 8748-8763.

[24] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB/OL]. (2023-02-27)[2024-03-30]. https://arxiv.org/abs/2302.13971.
[25] KAYSER M, CAMBURU O M, SALEWSKI L, et al. E-ViL: a dataset and benchmark for natural language explanations in vision-language tasks [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 1224-1234.
[26] YANG Z Y, GAN Z, WANG J F, et al. An empirical study of GPT-3 for few-shot knowledge-based VQA[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(3): 3081-3089.
[27] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and clustering [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 815-823.
[28] XIE N, LAI F, DORAN D, et al. Visual entailment: a novel task for fine-grained image understanding[EB/OL]. (2019-01-20)[2024-03-30]. https://arxiv.org/abs/1901.06706.
[29] PLUMMER B A, WANG L W, CERVANTES C M, et al. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models[C]//2015 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2015: 2641-2649.
[30] DO V, CAMBURU O M, AKATA Z, et al. E-SNLI-VE: corrected visual-textual entailment with natural language explanations[EB/OL]. (2021-08-19)[2024-03-15]. https://arxiv.org/abs/2004.03744.
[31] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments [C]// Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Stroudsburg: ACL, 2005: 65-72.
[32] ZHANG T Y, KISHORE V, WU F, et al. BERTScore: evaluating text generation with BERT[EB/OL]. (2020-01-24)[2024-04-23]. https://arxiv.org/abs/1904.09675.
[33] MARASOVIĆ A, BHAGAVATULA C, PARK J S, et al. Natural language rationales with full-stack visual reasoning: from pixels to semantic frames to commonsense graphs[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: ACL, 2020: 2810-2829.
[34] SAMMANI F, MUKHERJEE T, DELIGIANNIS N. NLX-GPT: a model for natural language explanations in vision and vision-language tasks [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 8312-8322.

基于特定指令驱动的上下文感知视觉蕴含

韩宇凤[1,2]，郝矿荣[1,2*]，唐雪嵩[1,2]，隗　兵[1,2]
1. 东华大学 信息科学与技术学院，上海 201620
2. 东华大学 数字化纺织服装技术教育部工程研究中心，上海 201620

摘　要：视觉蕴含是多模态视觉推理中的一个典型任务，当前的方法经常利用大型语言模型（large language model，LLM）作为知识库来协助回答问题。这些方法在很大程度上依赖于文本模态，而文本模态本质上无法捕获图像中包含的全部信息。为此，作者提出了一个上下文感知视觉蕴含（context-aware visual entailment，CAVE）模型。该模型引入了一种新的聚合模块，用于从图像中提取高级语义特征，将低级语义图像特征聚合为格式类似于文本标记的高级视觉标记，作为 LLM 的输入。CAVE 模型弥补了图像信息的损失，并更有效地将图像信息与文本理解相结合。同时，CAVE 模型采用了一种新的基于指令微调和上下文学习的输入格式和训练方法，其目的在于最大化 LLM 固有的逻辑推理潜能。在 E-SNLI-VE 数据集上的实验结果表明，CAVE 模型表现出色。

关键词：视觉蕴含；文本-视觉融合；指令微调；上下文学习